



**University of  
Nottingham**

UK | CHINA | MALAYSIA

# Visualisation of Electronic Health Records

Submitted to the University of Nottingham, in fulfillment of the requirements  
for the Degree of Doctor of Philosophy.

**Qiru Wang**  
**20263716**

**Supervised by Robert S. Laramee**

School of Computer Science  
University of Nottingham

Signature \_\_\_\_\_

Date \_\_\_\_ / \_\_\_\_ / \_\_\_\_

I hereby declare that I have all necessary rights and consents to publicly  
distribute this dissertation via the University of Nottingham's e-dissertation  
archive.



# Abstract

Electronic Health Records (EHRs) contain complex, multidimensional data, making effective visualisation essential for clinicians, researchers, and policymakers. This thesis investigates the challenges in EHR visualisation (EHR Vis) and explores novel solutions to improve data interpretation and decision-making.

We begin with a systematic review of state-of-the-art EHR Vis techniques, classifying existing methods, identifying limitations, and outlining opportunities for innovation. Based on this foundation, we develop three novel visualisation tools in collaboration with domain experts.

In [Chapter 3](#), we present a letter-space visualisation tool to support the exploration of unstructured clinical text describing epilepsy patients in a structured manner. In [Chapter 4](#), we introduce a novel hybrid cartogram layout algorithm that enhances the legibility, readability, and overall accuracy of EHR data visualised through Demers Cartogram. In [Chapter 5](#), we describe a hierarchical visualisation technique designed to efficiently extract, organise, and present event-based temporal data in long-term patient records.

A special chapter is presented in [Chapter 6](#), where we present EnsembleDashVis, a visualisation dashboard developed in collaboration with over 40 experts during the COVID-19 pandemic. This project demonstrates the role of visualisation in supporting large-scale, multidisciplinary emergency response efforts.

Through these contributions, this thesis advances the field of EHR Vis by addressing critical usability challenges, proposing scalable solutions, and demonstrating the value of visualisation in clinical and epidemiological contexts. We conclude with a discussion on the broader implications of our findings and propose future research directions to further enhance EHR Vis methodologies.





## Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Robert S. Laramee, for his continuous support and guidance throughout my Ph.D. and Master's studies. His expertise and passion for research have been truly inspiring, and our collaborations, including navigating the COVID-19 pandemic together, have been invaluable.

To all the domain experts, colleagues, and collaborators who contributed their time, insights, and feedback, thank you. Your contributions have enriched this research in ways I could not have achieved alone.

Finally, I would like to thank family, friends, and Carrot the cat for their unwavering support, encouragement, and sometimes distractions (mostly Carrot). Special thanks to my partner for her love and patience throughout this seemingly never-ending Ph.D. journey.

This Ph.D. received support from multiple grants funded by the Engineering and Physical Sciences Research Council (EPSRC), grant number EP/S010238/1, EP/S010238/2 and EP/V054236/1. EPSRC is a British Research Council that provides government funding for grants to undertake research and postgraduate degrees in engineering and the physical sciences.





## Contributions

This thesis is based on the following publications:

1. **Wang, Q.**, Laramée, R. S., Lacey, A., & Pickrell, W. O. (2021). LetterVis: A letter-space view of clinic letters. *The Visual Computer*, 37(9–11), 2643–2656. <https://doi.org/10.1007/s00371-021-02171-w> [298]
2. **Wang, Q.**, & Laramée, R. S. (2022). EHR STAR: The State-Of-the-Art in Interactive EHR Vis. *Computer Graphics Forum*, 41(1), 69–105. <https://doi.org/10.1111/cgf.14424> [311]
3. **Wang, Q.**, Borgo, R., & Laramée, R. S. (2024). EnsembleDashVis Views and Volunteers – A Retrospective and Early History. In M. Bassanello, R. Geppini, X.-N. Li, & A. Matecki (Eds.), *New Community Health Models*. IntechOpen. <https://doi.org/10.5772/intechopen.115029> [318]
4. **Wang, Q.**, Xu, K., & Laramée, R. S. (2024). Demers cartogram with rivers. *Visual Informatics*. <https://doi.org/10.1016/j.visinf.2024.09.003> [319]

The following manuscripts are under review at the time this thesis was submitted:

1. **Wang, Q.**, Bartolomeo, S. D., Dunne, C., Laramée, R. S., Litchfield, I., Weber, P., & Xu, K. (2024). Time Series Maps: Hierarchical Visualization of Blood Glucose Time Series Data. Manuscript submitted to *BMC Bioinformatics*.

The following papers were published during the Ph.D. period as the result of collaborative work on relevant research topics:

1. Chen, J., Ling, M., Li, R., Isenberg, P., Isenberg, T., Sedlmair, M., Moller, T., Laramée, Robert, R., Shen, H.-W., Wünsche, K., & **Wang, Q.** (2020). (2020). IEEE Vis Figures and Tables Image Dataset. *IEEE DataPort*. <https://doi.org/10.21227/4HY6-VH52> [253]
2. Rees, D., **Wang, Q.**, & Laramée, R. S. (2020). The industry engagement ladder. *Journal of Industry-University Collaboration*, 2(3), 125–139. <https://doi.org/10.1108/JIUC-02-2020-0001> [272]
3. 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., Laramée, R. S., Nguyen, P. H., Reeve, R., Roberts, J. C., Vidal, F., **Wang, Q.**, Wood, J., Xu, K. (2020). RAMPVIS: Towards a New Methodology

for Developing Visualisation Capabilities for Large-scale Emergency Responses.

<https://doi.org/10.48550/ARXIV.2012.04757> [254]

4. Liu, X., Alharbi, M., Best, J., Chen, J., Diehl, A., Firat, E., Rees, D., **Wang, Q.**, & Laramée, R. S. (2021). Visualization Resources: A Starting Point. The 25th International Conference on Information Visualization, 160–169. <https://doi.org/10.1109/IV53921.2021.00034> [293]
5. Chen, J., Ling, M., Li, R., Isenberg, P., Isenberg, T., Sedlmair, M., Moller, T., Laramée, R. S., Shen, H.-W., Wunsche, K., & **Wang, Q.** (2021). VIS30K: A Collection of Figures and Tables From IEEE Visualization Conference Publications. IEEE Transactions on Visualization and Computer Graphics, 27(9), 3826–3833. <https://doi.org/10.1109/TVCG.2021.3054916> [289]
6. Chen, M., Abdul-Rahman, A., Archambault, D., Dykes, J., Ritsos, P. D., Slingsby, A., Torsney-Weir, T., Turkay, C., Bach, B., Borgo, R., Brett, A., Fang, H., Jianu, R., Khan, S., Laramée, R. S., Matthews, L., Nguyen, P. H., Reeve, R., Roberts, J. C., Vidal, F. P., **Wang, Q.**, Wood, J., Xu, K. (2022). RAMPVIS: Answering the challenges of building visualisation capabilities for large-scale emergency responses. Epidemics, 39, 100569. <https://doi.org/10.1016/j.epidem.2022.100569> [301]
7. Dykes, J., Abdul-Rahman, A., Archambault, D., Bach, B., Borgo, R., Chen, M., Enright, J., Fang, H., Firat, E. E., Freeman, E., Gönen, T., Harris, C., Jianu, R., John, N. W., Khan, S., Lahiff, A., Laramée, R. S., Matthews, L., Mohr, S., ... **Wang, Q.**, Xu, K. (2022). Visualization for epidemiological modelling: Challenges, solutions, reflections and recommendations. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 380(2233), 20210299. <https://doi.org/10.1098/rsta.2021.0299> [302]

For existing and future publications, please visit Google Scholar <https://scholar.google.com/citations?user=sWXPmkQAAAAJ&hl=en> and ORCID <https://orcid.org/0000-0003-3397-308X>.

# Contents

<b>List of Tables</b>	<b>xi</b>
<b>List of Figures</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Data Visualisation	2
1.1.1 Information Visualisation	3
1.1.2 Visual Analytics	4
1.1.3 Text Visualisation	5
1.1.4 Geospatial Visualisation	5
1.1.5 Electronic Health Records Visualisation	7
1.2 Challenges	7
1.3 Research Methodology	9
1.4 Contributions	10
1.5 Thesis Structure	11
<b>2 EHR STAR: The State-Of-the-Art in Interactive EHR Vis</b>	<b>13</b>
2.1 Introduction and Motivation	15
2.1.1 Survey Challenges	16
2.1.2 Literature Search Methodology	17
2.1.3 Survey Scope	22
2.1.4 Background and Terminology	23
2.2 Literature Classification	27
2.2.1 Multidisciplinary Research Themes	27
2.2.2 Adopting a Medical Terminology Standard	27
2.3 Related Work	29
2.3.1 Related Work with an EHR Focus	30
2.3.2 Related Work with a PopHR Focus	35
2.4 EHR Vis	36
2.4.1 Machine Learning	37
2.4.2 Natural Language Processing	40
2.4.3 Event Sequence Simplification	42
2.4.4 Visual Analytics and Comparison	45
2.4.5 Visual Analytics with Clustering and Others	48
2.4.6 PopHR Vis and Geospatial Visualisation	49
2.5 Evaluation	52
2.6 Open Access Healthcare Data	55
2.6.1 Healthcare Data Challenges	55

2.6.2	Healthcare Data Search Methodology . . . . .	56
2.6.3	Healthcare Data Scope . . . . .	56
2.6.4	Healthcare Data Sources Classification . . . . .	57
2.6.5	Open Access Healthcare Data Sources . . . . .	58
2.7	Future Research Challenges and Discussion . . . . .	63
2.8	Conclusions . . . . .	69
<b>3</b>	<b>LetterVis: a Letter-space View of Clinic Letters</b>	<b>71</b>
3.1	Introduction . . . . .	73
3.1.1	Motivation . . . . .	73
3.1.2	Contribution . . . . .	74
3.2	Related Work . . . . .	75
3.3	Data Description . . . . .	77
3.4	Application Design Methodology . . . . .	78
3.4.1	Informing the Initial Design . . . . .	78
3.4.2	Informing Further Software Iterations . . . . .	79
3.4.3	User Requirements and Design Goals . . . . .	79
3.4.4	Tasks . . . . .	80
3.5	LetterVis for Visualisation of EHR Letters . . . . .	81
3.5.1	Initial Prototype . . . . .	82
3.5.2	Three Levels of Letter Abstraction . . . . .	84
3.5.3	Advanced Visual Filtering and Selection . . . . .	86
3.6	Evaluation . . . . .	90
3.6.1	Case Studies . . . . .	90
3.6.2	Domain Expert Feedback . . . . .	94
3.6.3	Domain Expert Review . . . . .	96
3.7	Limitations and Future Work . . . . .	98
3.8	Conclusions . . . . .	98
3.9	Appendix . . . . .	100
3.9.1	List of Expert Sessions . . . . .	100
3.9.2	List of Interview Questions . . . . .	100
<b>4</b>	<b>Demers Cartogram with Rivers</b>	<b>101</b>
4.1	Introduction . . . . .	103
4.2	Related Work . . . . .	104
4.3	Data Description . . . . .	108
4.3.1	Choropleth Shapefile . . . . .	108
4.3.2	River Shapefiles . . . . .	109
4.3.3	EHR Data . . . . .	110
4.4	Demers Cartogram with Rivers . . . . .	110
4.4.1	Initialisation with Rivers . . . . .	111
4.4.2	Node Layout and Overlap Removal . . . . .	111
4.4.3	River Intersection Testing . . . . .	113
4.4.4	Translating Rivers . . . . .	114
4.4.5	Process Stalemates . . . . .	116
4.4.6	Terminating the Algorithm . . . . .	118
4.4.7	User Options . . . . .	119
4.5	User Evaluation . . . . .	120
4.5.1	Study Hypothesis . . . . .	121

4.5.2	User Study Variables . . . . .	121
4.5.3	User Study Design . . . . .	122
4.5.4	User Study Analysis . . . . .	125
4.6	Limitations and Future Work . . . . .	130
4.6.1	Experimental Design Confounds . . . . .	130
4.6.2	Colormap choice . . . . .	131
4.6.3	Overlap removal algorithm choice . . . . .	131
4.6.4	Generalisability . . . . .	132
4.6.5	Improved User Study . . . . .	132
4.7	Conclusions . . . . .	132
4.8	Appendix . . . . .	133
4.8.1	Preprocessing Shapefiles . . . . .	133
4.8.2	List of Likert Scale Questions . . . . .	139
<b>5</b>	<b>Time Series Map</b>	<b>141</b>
5.1	Introduction . . . . .	143
5.1.1	Terminology . . . . .	144
5.1.2	Contributions . . . . .	144
5.2	Related Work . . . . .	145
5.2.1	Event Sequence Visualisation . . . . .	145
5.2.2	Co-occurrence Pattern Visualisation . . . . .	147
5.2.3	Clustering and Classification . . . . .	147
5.2.4	Visualisation of Blood Glucose Data . . . . .	148
5.3	Event Specification and Extraction . . . . .	149
5.3.1	Natural Language Event Specification . . . . .	149
5.3.2	From Natural Language to a Technical Event Specification . . . . .	152
5.3.3	Event Extraction . . . . .	154
5.4	Time Series Map . . . . .	155
5.4.1	Hierarchy Construction . . . . .	155
5.4.2	Time Series Map View . . . . .	156
5.4.3	Color and Threshold . . . . .	157
5.4.4	Other Views and Interactions . . . . .	160
5.5	Evaluation . . . . .	161
5.5.1	Data sets . . . . .	162
5.5.2	Case Study . . . . .	162
5.5.3	Health Data Experts Interviews . . . . .	164
5.6	Limitations and Future Work . . . . .	171
5.7	Conclusions . . . . .	171
<b>6</b>	<b>EnsembleDashVis Views and Volunteers - A Retrospective and Early History</b>	<b>173</b>
6.1	Introduction and Motivation . . . . .	175
6.2	Background and Related Work . . . . .	177
6.2.1	VIS for Emergency Response . . . . .	178
6.2.2	VIS for COVID-19 Data Modelling . . . . .	178
6.3	Data Description . . . . .	180
6.4	EnsembleDashVis . . . . .	182
6.4.1	An Unconventional Software Development Cycle . . . . .	183
6.4.2	Technology and Design . . . . .	184




6.4.3	Interaction	185
6.4.4	Meetings and Milestones	186
6.5	Domain Expert Feedback	194
6.5.1	Summary of Feedback	194
6.5.2	Detailed Feedback	195
6.6	Limitations	197
6.7	Conclusions	199
<b>7</b>	<b>Conclusion</b>	<b>201</b>
7.1	Key Contributions and Findings	202
7.1.1	Summary of Chapters	202
7.2	Future Work and Challenges	204
	<b>Bibliography</b>	<b>205</b>
	<b>Appendices</b>	
	<b>Appendix A Damon Berridge</b>	<b>241</b>
	<b>Appendix B List of Domain Experts</b>	<b>243</b>
B.1	Alfie Abdul-Rahman	243
B.2	Sara Di Bartolomeo	243
B.3	Rita Borgo	243
B.4	Peter Challenor	244
B.5	Min Chen	244
B.6	Alena Denisova	244
B.7	Cody Dunne	244
B.8	Arron Lacey	244
B.9	Ian Litchfield	245
B.10	Owen Pickrell	245
B.11	Panagiotis D. Ritsos	245
B.12	Benjamin Swallow	245
B.13	Thomas Torsney-Weir	246
B.14	Cagatay Turkey	246
B.15	Samantha Turner	246
B.16	Ian Vernon	246
B.17	Franck P. Vidal	246
B.18	Phil Weber	247
B.19	Kai Xu	247
	<b>Appendix C List of Data sets Used</b>	<b>249</b>
	<b>Appendix D List of Data Set Access Applications</b>	<b>251</b>
	<b>Appendix E List of Videos</b>	<b>253</b>



# List of Tables

2.1	Conferences and Journals (both Visualisation and NonVisualisation venues) used for discovering literature and the number of papers found. . . . .	18
2.2	Keyword combinations used for discovering EHR Vis literature. . . . .	19
2.3	<b>UMLS table:</b> 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 [32]. 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 visualised on the x-axis. <b>Green</b> highlights context papers included in this STAR. . . . .	21
2.4	<b>Terminology table:</b> 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 2.3. This table indicates a mixture of terms is used throughout the literature. We clarify the terminology in Section 2.1.4. <b>Green</b> highlights context papers. . . . .	25
2.5	<b>Focus papers:</b> Y-axis, common Focus papers from previous survey papers, ordered by the year of publication. X-axis, <b>E</b> indicates an EHR focused survey and <b>P</b> indicates a PopHR focused survey. We can see that some of previously published EHR Vis papers are common to multiple surveys. . . . .	29
2.6	<b>Context papers:</b> Y-axis, overlapping context papers from previous survey papers, ordered by the year of publication. X-axis, <b>E</b> indicates an EHR focused survey and <b>P</b> indicates a PopHR focused survey. We can observe that the 2013 survey by Rind et al. [126] has some thematic overlap with this one. . . . .	30
2.7	<b>Out of scope papers:</b> Y-axis, out of scope papers from previous survey papers, ordered by the year of publication, with the <b>exclusion criteria</b> described in Section 2.1.3: (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, <b>E</b> indicates an EHR-focused survey and <b>P</b> indicates a PopHR-focused survey. . . . .	31
2.8	<b>Overview of EHR Vis techniques:</b> Ordered by the publication year. The x-axis is mapped to the re-occurring research themes we extracted from the literature. <b>Red</b> highlights the primary theme, <b>Grey</b> highlights the secondary theme, and <b>Green</b> highlights context papers. . . . .	37
2.9	An overview table of ML topics discussed in the literature described in Section 2.4.1. Papers with ML as a secondary theme are highlighted in <b>Green</b> . . . . .	38

2.10	An overview table of NLP approaches adopted by the literature described in Section 2.4.2. Papers with NLP as a secondary theme are highlighted in <b>Green</b> .	40
2.11	An overview table of event types in the literature described in Section 2.4.3. Papers with ESS as a secondary theme are highlighted in <b>Green</b> .	44
2.12	An overview table of comparative designs adopted by the literature described in Section 2.4.4. The x-axis is mapped to the comparative design categorisation by Gleicher et al. [90]. Papers with Comparison as a secondary theme are highlighted in <b>Green</b> .	46
2.13	An overview table of clustering dimensions used in the literature described in Section 2.4.5.	48
2.14	An overview table of geospatial regions covered in the literature described in Section 2.4.6.	50
2.15	<b>Evaluation table:</b> 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. <b>Green</b> highlights context papers.	54
2.16	Keyword combinations used for discovering relevant healthcare data.	56
2.17	<b>Data source table:</b> Data sources ordered by the year of establishment. See the detailed description of focus data sources in Section 2.6.5. <b>Green</b> highlights context data sources. <sup>C</sup> Contains COVID-19 data. <sup>†</sup> Registration required for open access. <sup>‡</sup> Partially open access. <sup>‡‡</sup> Free access for project collaborators, paid access for noncollaborators. <sup>¶</sup> Free access for project collaborators, no access for noncollaborators. <sup>††</sup> Data is not archived in English.	58
2.18	<b>Challenge table:</b> 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. <b>Green</b> highlights context papers. We use 1-2 words to represent these challenges in the table header, and describe them in detail in Section 2.7.	65
2.19	Visualisation techniques applied in the literature. We follow the classification of visualisation techniques by Keim[26], and categorise bar chart, line chart and pie chart as standard 2d display. <b>Green</b> highlights context papers. ● indicates the technique is applied in the literature. ● indicates a customised variant of the technique is applied in the literature.	68
3.1	12 text data categories extracted from our collection of letters, with a brief description and the total number of items extracted for each category.	78
3.2	Boolean operators <b>AND</b> , <b>OR</b> and <b>NOT</b> are supported with (and ) for grouping. Categories are in all capitals and can be used in the query to highlight all items belonging to the category. The colour legend in Figure 3.3A left shows the list of available categories.	88

3.3	The resulting letters after searching and filtering based on pregnancy and AED. Letters are classified into five categories: 1) Patients with pregnancy planned or potentially planned. 2) Patients with no pregnancy planned. 3) Patients with pregnancy completed. 4) Letters containing keywords ‘pregnancy’ and ‘pregnant’ but are irrelevant to the patient’s condition, such as describing the patient’s own birth. 5) A special case is described in detail in Case Study 3 - AEDs and Pregnancy. <b>Green</b> highlights letters with Sodium Valproate prescribed. . . . .	95
3.4	Summary of domain expert sessions (both interviews and feedback), conducted in person and virtually due to the COVID-19 pandemic restrictions. . . . .	100
4.1	Related work with noncontiguous cartogram-based visualisations. <b>Cartogram type</b> is the type of cartogram used. <b>Geographic region</b> is the geographic region depicted by the cartogram. <b>Number of nodes</b> is the number of nodes (representing geographic enumeration units) depicted in the cartogram. . . . .	105
4.2	 Trade-off between dimensions.  Dimension sacrificed in order to improve  target dimension’s accuracy. . . . .	106
4.3	The file size is reduced by 88.5% from the original size. . . . .	109
4.4	The table compares various statistical measures of response times in milliseconds (ms), showing the original values versus those after removing outliers. Outliers were defined as values above $Q3 (12,839 \text{ ms}) + 1.5 \times IQR (24,716 \text{ ms})$ . . . . .	126
4.5	The table presents the accuracy and mean response times (in milliseconds) of four different conditions, after removing outliers. . . . .	126
5.1	The table presents a list of event sequence visualisation research with the scope of Electronic Health Records (EHR) published over the last decade. We categorise the papers based on their <i>Visual Representation</i> , <i>Number of Events Rendered</i> , and <i>Subject of Data (UMLS Code)</i> , where $n_c$ denotes the number of groups and clusters rendered. . . . .	146
6.1	16 input parameters for the ABC-SMC inference model. As constant parameters such as $K$ and $rrd$ do not affect the simulation results, they are not rendered in our visual designs. . . . .	180
6.2	13 output parameters from the simulation performed by the ABC-SMC inference model. . . . .	181
6.3	The table shows the list of meetings held throughout the entire volunteering period, detailing each meeting’s date, the attendees, and the milestones accomplished. . . . .	187
C.1	A list of data sets explored in this Ph.D. . . . .	249
D.1	A list of EHR data set access applications during this Ph.D. This table demonstrates the difficulty in accessing EHR data sets for research purposes. . . . .	251
E.1	A collection of videos showcasing the research conducted during this Ph.D. . . . .	253



# List of Figures

1.1	The map of England and Wales by John Adams in 1677. Figure reproduced from Götzfried Antique Maps [348]. . . . .	2
1.2	The exports and imports to and from Denmark and Norway in England between 1700 and 1780. Figure reproduced from Playfair [1]. . . . .	3
1.3	Estimated 10-year migrant flows during 2000-10 between regions. Figure reproduced from Abel [203]. . . . .	4
1.4	An example of a treemap visualising London property transactions between 2000 and 2008. Figure reproduced from Slingsby et al. [69]. . . . .	4
1.5	TransVis facilitates the comparison of 38 translations of Othello over a span of 244 years. Figure reproduced from Alharbi et al. [300]. . . . .	6
1.6	An example of a Mosaic cartogram depicting the US election results in 2012. Figure reproduced from Cano et al. [146]. . . . .	6
1.7	An example of a choropleth map depicting the output areas of Wales in 2016. Figure reproduced from McNabb et al. [213]. . . . .	7
1.8	A Venn diagram illustrates the relationships between EHR Vis and other subfields of Data Visualisation. . . . .	8
2.1	The various subdomains integrated in the UMLS Terminology. Figure reproduced from Abel [32]. . . . .	28
2.2	PhenoLines [209] 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. Figure reproduced from Glueck et al. Glueck et al. [209]. . . . .	39
2.3	EventFlow [122] visualising the original Long-Acting $\beta$ -Agonists data set on the left, and the simplified data set on the right. The number of visual elements is reduced from 2,700 to 492. Figure reproduced from Monroe et al. [122]. . . . .	43
2.4	PhenoStacks [193] 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 visualises 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. Figure reproduced from Glueck et al. [193]. . . . .	45
3.1	The multilevel typology of abstract visualisation tasks by Brehmer and Munzner. Figure reproduced from Brehmer and Munzner [118]. . . . .	80

3.2	An illustration of the colour legend for 12 numerical categories. Clicking on categories will render the corresponding centroids and individual samples in the global, thumbnail and focus views as context (in greyscale). Each data category is followed by the number of matches found in the uploaded letters. . . . .	82
3.3	This figure shows an overview of LetterVis. We also provide a detailed description of visual designs and elements in their respective sections. (A) shows the user options for searching, rendering, and sorting (Section 3.5.3). (B) illustrates the matrix view based on AED co-occurrences in the data set, user-chosen cells are highlighted with a sequence number, which corresponds to their position in the history of queries (Section 3.5.3). (C) The drug chain layout returned by the queries, user-chosen AEDs are shown in colour (Section 3.5.3). (D) depicts an overview of all super-imposed letters and their search result centroids in the data set (Section 3.5.2). (E) shows individual thumbnails of letters with local search result centroids joined by edges (Section 3.5.2). (F) contains a detailed view of the letter in focus, lines without data-of-interest are collapsed by default (Section 3.5.2). . . . .	83
3.4	Illustrations of letter alignment in the drug chain view. A) The initial layout of the drug chain view. B) Letters are aligned via clicking on a base AED, Lamotrigine, highlighted with a red border in the first letter. Letters without the AED are shown in context with reduced opacity. C) Letters are sorted by alignment, with context letters being shifted to the end of the queue. Hovering over any block will display a tooltip containing the letter title and AED name. D) All AEDs are put in focus mode (in color) via a toggle. Chains are then sorted by the number of AEDs. E) Ethosuximide, a rare prescription in the data set, is selected as the base AED for alignment. F) Chains are sorted by gender, with a horizontal grey bar as the separator. The same subset of letters is used in this figure. . . . .	87
3.5	An illustration of centroid exploration in the global and thumbnail views. The top shows the default search and rendering when loading 30 letters. By default, the global view searches and renders 12 text data categories in focus and renders individual data samples in greyscale, a classical focus+context approach. A thumbnail view is presented with identical default rendering options for individual letters. The bottom shows the edges connecting the user-chosen (focus) centroid, <i>DRUG</i> , and individual samples in both views. Other centroids and individual samples are rendered as context. Edges can also be hidden as an option. . . . .	89
3.6	A screenshot of one letter’s original view, with personally identifiable information redacted. Search query terms are highlighted in their corresponding colours, as described in Section 3.5.2. . . . .	91
3.7	We execute the query ‘(pregnant <i>OR</i> pregnancy) <i>AND</i> DRUG’ and sort letters by gender to focus on pregnancy. The thumbnail view shows two male patients separated from the rest by a grey bar. High-risk cases discovered in Case Study 3 are highlighted with a red line. . . . .	94
4.1	A map of 135 CCGs in England as of 2020, obtained from the Open Geography Portalx [352] with EPSG:4326 (WGS84 - World Geodetic System) as the Coordinate Reference System (CRS). . . . .	108

4.2	An overview of our hybrid layout algorithm incorporating rivers. See also Algorithm 1 in Section 4.4 for more detail. See Figure 4.4 for the logic of processing a stalemate. For illustration purposes, we show the rendered views alongside the logical views representing the actual computation, and use the same size for all squares for clarity. . . . .	111
4.3	The resolution of rivers can be dynamically adjusted by the user. (A) shows River Thames at its original resolution with 10,170 edges. (B) shows the river at a reduced resolution of 49 edges. We further smooth the river by removing 19 vertices in dense areas, as shown in (C). The reduced resolution preserves the majority of River Thames' original shape and improves the performance of our river intersection tests. . . . .	112
4.4	A flowchart illustration of stalemate processing. See Section 4.4.5, Figure 4.5, and refer to our video demonstration (from 1:27 to 1:46) for details. . . . .	117
4.5	A stalemate: when a node's translation path $\overrightarrow{nn_t}$ (iteration 1) intersects a river $w$ times. The node is translated back to its previous position (iteration 2). A stalemate can occur when the area is congested and the node cannot translate to a new position without intersecting a river. . .	118
4.6	A stalemate occurs when a node's translation path $\overrightarrow{nn_t}$ intersects a river for $w$ times, as shown in (A). To address this, we derive a corridor (orange rectangle in (E)) based on $\mathbf{n}$ and $\mathbf{n}_t$ . All nodes within the corridor are translated based on $\overrightarrow{n_t n}$ , such that $\overrightarrow{nn_t} = \overrightarrow{n_{in} n_{in_t}}$ . For clarity in the illustration, we place nodes sparsely in this figure. . . . .	119
4.7	A screenshot of the user interface. User options are provided to adjust the terminating conditions (size and error), colour mapping, and visibility of nodes and rivers. Other options include the ability to control the overlap removal behaviour: rivers can be static or dynamic. See Section 4.4.7 for more on user options. In this figure, $\epsilon_{c_{max}} = 1.875\%$ , and an $\epsilon_t$ of 9.631% is eliminated. . . . .	121
4.8	A sample location task for participants. The left shows the choropleth map, and the right shows the corresponding cartogram. Both images show the three longest rivers in England, with the size and colour of nodes representing the prevalence of the selected disease. The target CCG blinks on the choropleth (shown in black), and participants are asked to identify this CCG on the cartogram. In this figure, $\epsilon_{c_{max}} = 1.875\%$ , and an $\epsilon_t$ of 6.667% is eliminated. . . . .	124
4.9	The stacked bar chart shows the user study participant responses of Likert Scale questions. . . . .	128
4.10	Due to colour and relative location, we believe the CCGs in the black circle are easier to locate. . . . .	131
4.11	QGIS interface, with River Trent, River Great Ouse, and River Thames (from top to bottom) imported. . . . .	134
4.12	QGIS interface, with all NHS CCGs imported. . . . .	135
4.13	QGIS interface, showing the unified CRS (OSGB36) for both layers. . .	136
4.14	QGIS interface, exporting all rivers using the OSGB36 CRS in GeoJSON. .	137
4.15	QGIS interface, exporting all NHS CCGs using the OSGB36 CRS in GeoJSON. . . . .	138
4.16	Mapshaper interface, merging all rivers with NHS CCGs into one layer, and export the merged layer in TopoJSON. . . . .	139

5.1	An overview of Time Series Maps to visualise blood glucose data. The hierarchical Time Series Map view on the left offers an overview of all the events, while the day-oriented view on the right provides a detailed view of the trends in blood glucose readings. The details-on-demand view at the bottom enables users to explore individual events and their corresponding blood glucose readings. In this figure, 8 months worth of data is rendered, which includes 68,000 events. . . . .	143
5.2	Six shape-based event categories reflecting blood glucose levels that may require attention. Category (A) depicts a monotonic sharp rise, while Category (B) illustrates a nonmonotonic sharp rise. Categories (C) and (D) represent monotonic and nonmonotonic sharp falls, respectively. Category (E) is characterised by a bell curve. Category (F) depicts a reverse bell curve. . . . .	150
5.3	This pair of 2D histogram matrices depict the distribution of both monotonic and nonmonotonic sharp rises and falls captured automatically using a duration indicated by the x-axis. The y-axis represents the change in readings as a percentage, and the x-axis represents the duration of events in minutes. Colour is simply mapped to number of events in each histogram bin. In this case, 8,513 events are automatically identified from the Marjorie data set [315]. . . . .	151
5.4	Similar to Figure 5.3, this 2D histogram matrix depicts both the number of monotonic and nonmonotonic sharp falls captured automatically. . .	152
5.5	A sample 2D histogram matrix depicts bell curve and reverse bell curve <i>events of interests</i> (EoIs) using a sliding window ( $k = 12$ ). The y-axis represents the percent change in glucose readings, calculated using the minimum and maximum values in each captured curve, while the x-axis is mapped to event duration. The distribution reveals that the subject's glucose readings show frequent but small spikes and drops, which could be used to infer the stability or variability of glucose levels over time. The majority of the events fall within the 0 - 20% change bracket. 17,679 curves are captured and rendered from the Marjorie data set [315]. . . .	152
5.6	This grouped bar chart depicts the distribution of threshold events in the Marjorie data set [315]. The y-axis represents the number of threshold events, and the x-axis represents the blood glucose thresholds. Ascent threshold events, e.g. 49 to 50, are shown in orange, and descent threshold events, e.g. 50 to 49, are shown in blue. . . . .	153
5.7	The hierarchy is derived from the unique shapes of the six shape-based event groups presented in Figure 5.2. Each group is further subdivided into a range of durations as indicated in Figure 5.8, which are mapped to the x-axis, while the changes in reading are mapped to the y-axis. . .	156
5.8	A Time Series Map constructed based on the hierarchy shown in Figure 5.7. It provides an overview of all the events, with the size of each parent node determined by the number of children in the corresponding event category. Each child node also contains a set of concentric rectangles that depict the frequency and distribution of threshold EoIs for that group. Superimposed over each child node is a curve provides a representative trend for the category, illustrating each group in Figure 5.2. This view enables quick identification of both the volume and characteristics of EoIs within each category. The colour scale is shown in Figure 5.10 bottom. . . . .	157



5.9	Important glucose target ranges, thresholds, and categories, proposed by the International Diabetes Center [117]. We use this as one of our colormap options, due to the importance of threshold-based events as guided by the health data experts we worked with. Figure reproduced from Bergenstal et al. [117]. . . . .	158
5.10	Two colormaps are available for representing glucose reading categories. Top: a colormap from the International Diabetes Center, as shown in Figure 5.9. Bottom: a sequential colormap derived from ColorBrewer [28], as suggested by a health data expert. . . . .	158
5.11	The figure illustrates treemap nodes with a concentric square colour design. Each node's colours are determined by the distribution of glucose readings per threshold category represented by the node. The numbers indicate the total number of EoIs for the child node. . . . .	159
5.12	The day-oriented view provides a more traditional depiction of blood glucose readings with each row representing one day divided into 24 hours. The x-axis represents the time of day, while the y-axis of each row represents the blood glucose readings in mg/dL. Each colored rectangle corresponds to one hour, with the horizontal curve within indicating the blood glucose readings. The gaps in the graph, where the rectangles are absent, are periods when readings are missing, indicating times when the continuous glucose monitor did not record. The red outline rectangles highlight patterns observed in case study 2. . . . .	161
5.13	The overview for case study 1. In this case study, we focus on finding unusual rises in glucose readings. . . . .	163
5.14	The overview for case study 3. In this case study, we focus on finding hypoglycemia, where glucose readings fall below 70 mg/dL. . . . .	165
5.15	Comparison of the Time Series Map view, with representative curves enabled (top) and disabled (bottom), using the colormap suggested by health data experts, which is derived from ColorBrewer [28]. . . . .	168
6.1	A timeline of the events between Mar 2020 and the end of our volunteer work on 19 May 2021. The upper section include policy changes during the time span, the lower section includes project developments and meetings. Milestone events are shown in red. . . . .	175
6.2	The organisation of researchers from the SCRC and RAMPVis. The SCRC modelling team is responsible for developing the epidemiological models leveraging different modelling techniques. The RAMPVis team provides visualisation support to the SCRC modelling team, by establishing four VIS volunteer teams who work on the actual development under the guidance of the RAMPVis team. . . . .	176
6.3	An illustration of the flow from the input parameters to the prediction results. 160 sets of input parameters are used to perform 1,000 simulation iterations, resulting in 160 sets of prediction results. . . . .	181

6.4	The overview of EnsembleDashVis. The dashboard consists of five views: (Figure 6.4A) a parallel coordinates plot for all input configurations, (Figure 6.4B) a table view with glyphs for all input configurations, (Figure 6.4C) a parallel coordinates plot with brushing to enable quick simulation outcomes filtering, (Figure 6.4D) a line chart for model predictions, and (Figure 6.4E) a scatterplot for Principal Component Analysis (PCA) outcomes. The views are coordinated with each other, enabling the modellers to observe relationships between input and outcome through interactions. . . . .	182
6.5	The structure of the actual code. Components are organised into separate files, with each file containing the code for a single view. Utilities contain the code for the data preprocessing and calculations. Data contains the metadata and preprocessed output by utilities. . . . .	185
6.6	The first visual design, a parallel coordinates plot depicting all 160 input configurations of the model, was completed on 5 Nov 2020. Each axis represents an input parameter, the y-axis represents the value of the parameter, and each polyline represents one input configuration. The table below shows the configuration details. . . . .	189
6.7	A line chart depicting the model outcomes. The x-axis of the chart corresponds to the number of days since the first date in the Scottish data set, while the y-axis represents the population. To differentiate between different population categories, a colormap was incorporated: <code>susceptible</code> , <code>exposed</code> , <code>hospitalised</code> , <code>recovered</code> , <code>death</code> , <code>asymptomatic</code> , and <code>symptomatic</code> . The focus+context technique is used here to highlight the outcome of the current configuration, while the grey lines represent other outcomes. On day 20, there is an unusual spike which was later identified as caused by an error in the model. . . . .	190
6.8	The table view depicting all 160 input parameter configurations. The view enables the modellers to sort parameter values and identify interesting configurations. Each row represents an input configuration, and each column represents an input parameter. Upon clicking on a row, the line chart in Figure 6.7 is updated to display the corresponding model outcomes. Clicking on the column header sorts the table by the parameter values. . . . .	191
6.9	A parallel coordinates plot depicting the model outcomes by age group 5. As requested by the modellers, each <code>blue</code> line represents one simulation outcome, and each coloured line represents the age group's mean. In addition, the dotted red line <code>...</code> represents the group's standard deviation, and the dashed red line <code>---</code> represents the mean of all groups. . . . .	193
6.10	A scatterplot depicting the PCA outcome from another VIS volunteer group, was added upon request by the modellers. Upon brushing, the selected configurations are highlighted in the table view in Figure 6.8. .	193

# Chapter 1

## Introduction

*“The greatest value of a picture is when it forces us to notice what we never expected to see.”*

– John W. Tukey, Mathematician (1915 - 2000)

### Contents

---

<b>1.1</b>	<b>Data Visualisation . . . . .</b>	<b>2</b>
1.1.1	Information Visualisation . . . . .	3
1.1.2	Visual Analytics . . . . .	4
1.1.3	Text Visualisation . . . . .	5
1.1.4	Geospatial Visualisation . . . . .	5
1.1.5	Electronic Health Records Visualisation . . . . .	7
<b>1.2</b>	<b>Challenges . . . . .</b>	<b>7</b>
<b>1.3</b>	<b>Research Methodology . . . . .</b>	<b>9</b>
<b>1.4</b>	<b>Contributions . . . . .</b>	<b>10</b>
<b>1.5</b>	<b>Thesis Structure . . . . .</b>	<b>11</b>

---



Figure 1.1: The map of England and Wales by John Adams in 1677. Figure reproduced from Götzfried Antique Maps [348].

## 1.1 Data Visualisation

The history of Data Visualisation dates back to the 17th century, when John Adams presented a survey of England and Wales with the measurement of boundary lines and distances between towns using latitude, angle, and direction [2]. The use of the combination of multiple elements such as points, lines, numbers, words, and symbols on top of a coordinate system is often regarded as the enlightenment of the early days Data Visualisation [22]. It was until the 19th century, when the common visual designs such as the line chart, bar chart, pie chart, and area chart, were invented by William Playfair in an attempt to replace boring tables of numerical data [1].

Modern Data Visualisation is capable of more than replacing statistical tables, and is defined as “the use of computer-supported, interactive, visual representations of data to amplify cognition” [13]. Over the past three decades, several subfields related to EHR Vis have been proposed, as identified in our literature review presented in Chapter 2. Key subfields include Information Visualisation, Visual Analytics, Text Visualisation, and Geospatial Visualisation. The rest of this section briefly describes these related subfields and examines their connection to the central topic of EHR Vis.

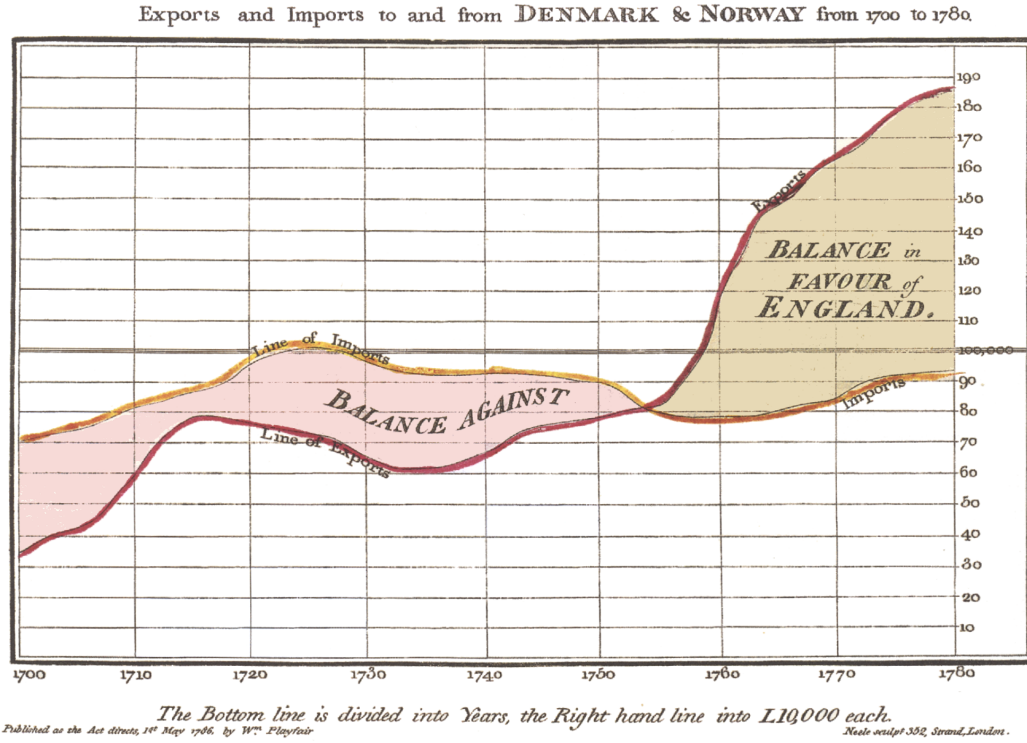


Figure 1.2: The exports and imports to and from Denmark and Norway in England between 1700 and 1780. Figure reproduced from Playfair [1].

### 1.1.1 Information Visualisation

Information Visualisation as a vibrant subfield, is defined by Stuart Card as “a set of technologies that use visual computing to amplify human cognition with abstract information.” [106]. Since abstract data has no inherent mapping to space, various visual representations of abstract data are proposed:

**Chord Diagram:** A chord diagram is a representation of the relationships between several categories of data. Each category is represented by a fragment of the outer circle, and the relationships between the categories are represented by the chords. The size of the chord is proportional to the number of items in the category. See Figure 1.3.

**Treemap:** A treemap is best used for the representation of hierarchical structures in the data [7]. Many variants have since been proposed to emphasise different data dimensions by adjusting the underlying layout algorithms [273], such as Nested Circles [31] and Balanced Partitioning [227]. See Figure 1.4.



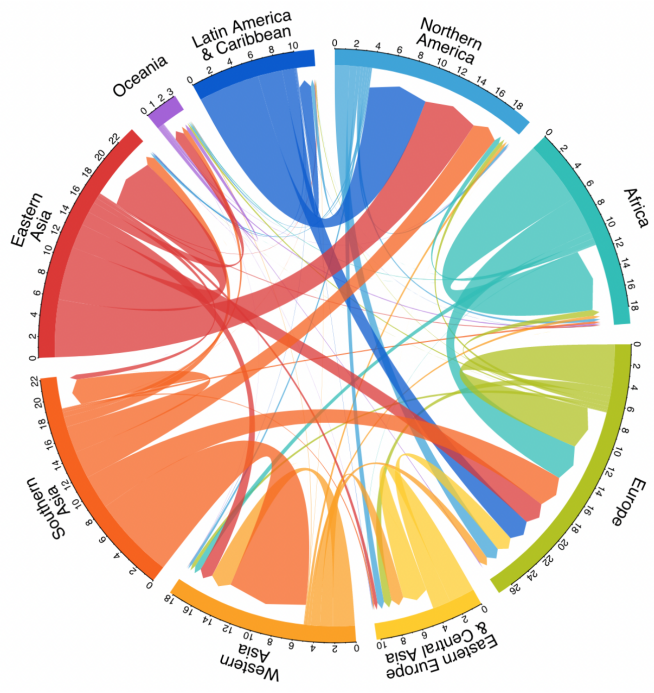


Figure 1.3: Estimated 10-year migrant flows during 2000-10 between regions. Figure reproduced from Abel [203].

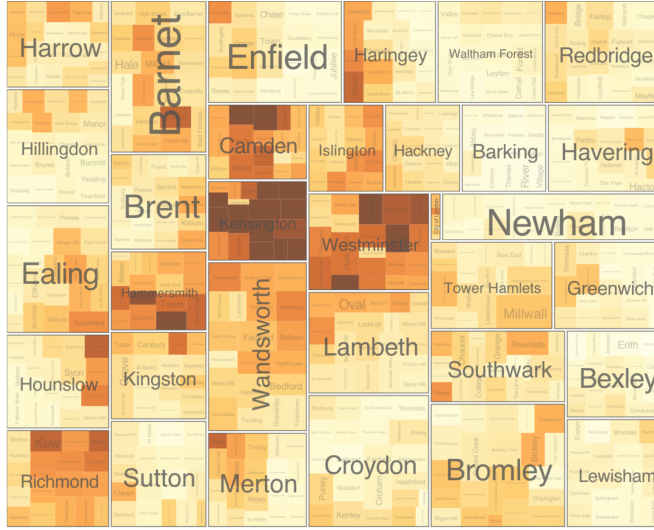


Figure 1.4: An example of a treemap visualising London property transactions between 2000 and 2008. Figure reproduced from Slingsby et al. [69].

### 1.1.2 Visual Analytics

Visual Analytics (VA) is often regarded as an extension of the fields of Scientific and Information Visualisation with technologies from other fields [35], thus the definition of VA varies from field to field. In the scope of this thesis, we follow the definition of VA as the combination of “automated analysis techniques with interactive visualisations for an effective understanding, reasoning, and decision-making on the basis of very large

and complex data sets.” [40].

As the amount of data being collected grows exponentially, we have entered an era of information overloading. VA aims to address this problem by extracting reliable knowledge from the massive heterogeneous data sets and communicating it to the user in an appropriate and interactive way [58]. Throughout this thesis, we present our research with novel interactive techniques to enable users to explore and comprehend the data that is being rendered.

### 1.1.3 Text Visualisation

As another popular subfield of Information Visualisation, Text Visualisation focuses on the representation of abstract text data and its relationships, to enable the discovery of actionable insights [159]. The use of visualisation techniques enables the analysis of massive text data that was previously difficult to perform, such as TransVis shown in Figure 1.5. In Chapter 3, we discuss the visualisation of unstructured text data, in the form of clinical letters, in detail.

### 1.1.4 Geospatial Visualisation

Geospatial Visualisation leverages geographic information to amplify legibility and recognisability. Geographic information is not limited to the geographic locations of the data, but can also include dimensions such as spatial distances, natural features (rivers and mountains), landmarks, and any important information related to the visualisation [295]. Here we describe some examples:

**Cartograms:** Cartograms are representations of geographical and abstract data based on a value-by-area mapping combining statistical and geographical information [66]. Each area is represented by a polygon such as a square (instead of the original shape). Each area varies in size and/or colour depending on the data shown on the map. See Figure 1.6 for an example. As an integral part of Chapter 4, we discuss cartograms in detail there.

**Choropleth Map:** A choropleth map is the representation of area-based data on a geospatial map [213]. Color is often used to encode values on a geographical area, visualising patterns and variations. See Figure 1.7.

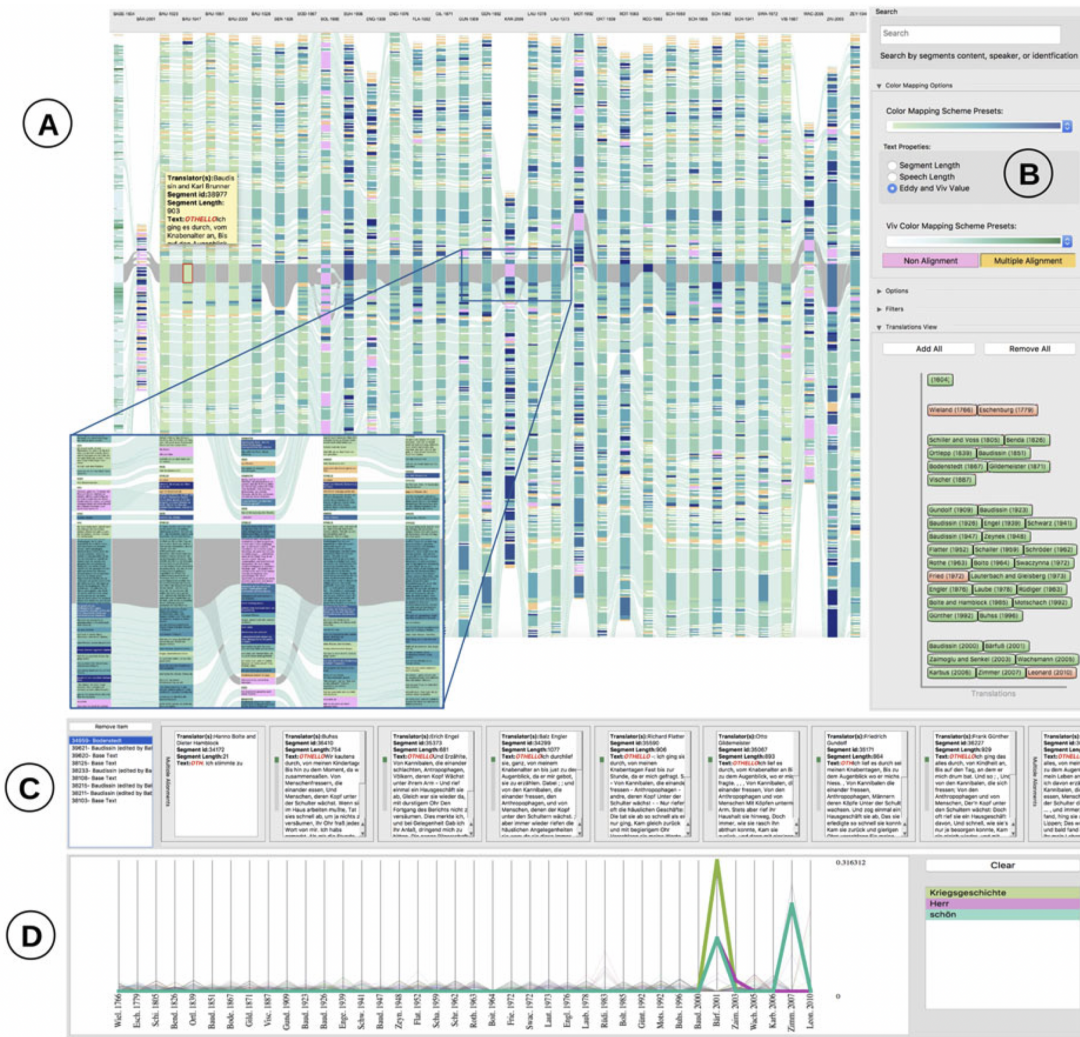


Figure 1.5: TransVis facilitates the comparison of 38 translations of Othello over a span of 244 years. Figure reproduced from Alharbi et al. [300].

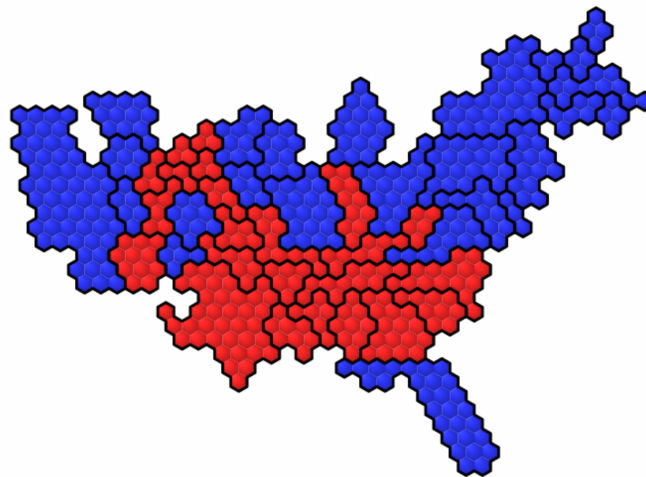


Figure 1.6: An example of a Mosaic cartogram depicting the US election results in 2012. Figure reproduced from Cano et al. [146].





Figure 1.7: An example of a choropleth map depicting the output areas of Wales in 2016. Figure reproduced from McNabb et al. [213].

### 1.1.5 Electronic Health Records Visualisation

To the best of our knowledge, there is no consensus on the definition of Electronic Health Records Visualisation (EHR Vis). In our state-of-the-art report described in [Chapter 2](#), we attempt to define EHR Vis as the visualisation of longitudinal collection of comprehensive patient medical information, maintained and shared by healthcare providers in machine-readable formats, and stored securely in an electronic system [311]. EHR Vis combines multiple subfields described earlier in this chapter, as shown in [Figure 1.8](#), in order to facilitate visual analytics and present the outcomes to policy-makers, clinicians, and patients.

As the core subject of this thesis, we break down EHR Vis into state-of-the-art in [Chapter 2](#), its combination with text visualisation in [Chapter 3](#), and geospatial visualisation in [Chapter 4](#).

## 1.2 Challenges

In this thesis, we consider the following as the main challenges:

1. **Diverse EHR Literature Sources:** As literature is spread across conferences and journals from different academic communities, it is challenging to understand the landscape of EHR Vis. We spent over a year investigating related work and

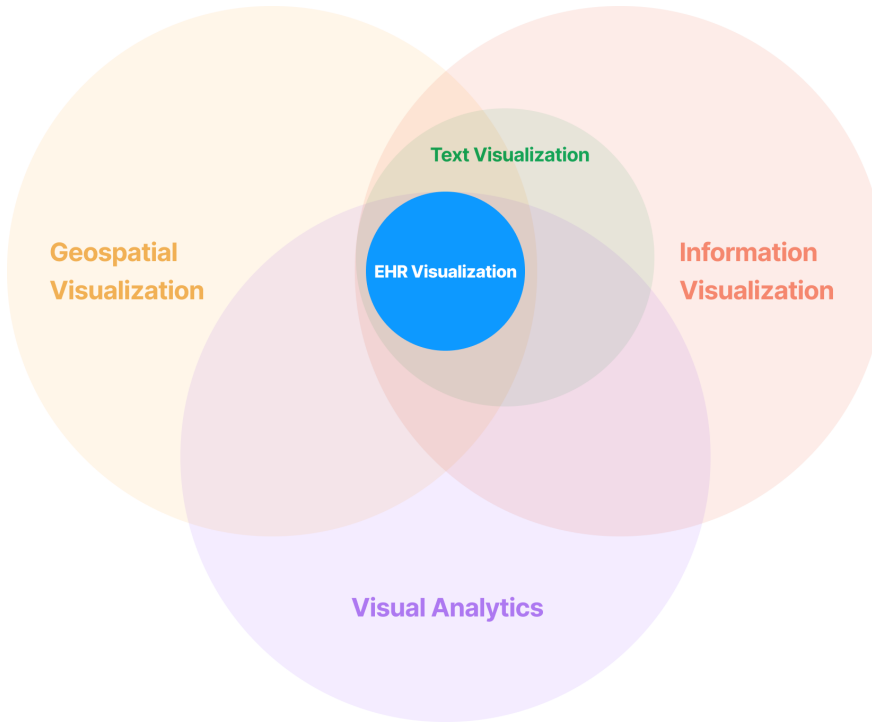


Figure 1.8: A Venn diagram illustrates the relationships between EHR Vis and other subfields of Data Visualisation.

the result is an up-to-date survey of EHR Vis literature, presented in [Chapter 2](#).

2. **EHR Data Acquisition:** The sensitive nature of EHR data poses challenges to researchers. Access to data often requires a lengthy application process. During the Ph.D. period, we attempted to apply for access to multiple EHR data sets and kept a record of processing time, see [Table D.1](#).
3. **EHR Data Scalability:** The size of an electronic healthcare data set is often huge. The rate of data growth exceeds the capacity of algorithms and software developed to visualise it [141]. In order to develop visual designs with a satisfactory level of performance, we need to carefully perform preprocessing steps to 1) minimise the data set size without losing valuable information, 2) improve the interactivity of the visual designs to facilitate the mantra of “overview first, details-on-demand”, furthermore 3) we employ state-of-the-art algorithms to ensure the performance of our visual designs when the data size grows.
4. **EHR Data Diversity:** Another challenge arises from the diversity of data types present in EHR data sets. These data sets often include structured data, such as numerical values and coded entries, alongside unstructured text, such as physician notes or diagnostic reports. Handling such heterogeneous data is

essential for creating effective visualisations, as different types of data require different preprocessing and visualisation techniques.

5. **EHR Vis Evaluation:** The evaluation of the visual design is critical to the success of EHR Vis. We need to carefully design the evaluation methods and work closely with domain experts.
6. **Others:** Other challenges stemming from nonresearch aspects. During the Ph.D. period, we witnessed the Covid-19 pandemic which severely disrupted our progress. However, we managed to contribute to the fight against the pandemic by developing EnsembleDashVis, as described in [Chapter 6](#).

## 1.3 Research Methodology

The challenges outlined in [Section 1.2](#) lay the groundwork for understanding the inherent complexities of visualising Electronic Health Records. These issues highlight the limitations of traditional visualisation methods and the necessity for innovative approaches. To address these challenges, we adopt the following research methodology. We begin by conducting a literature survey to understand the current state-of-the-art in EHR Vis. The survey also investigates and collects a list of open-access EHR data sets available for our research. The outcome of the survey enables us to have a clear overview of challenges and approaches that have already been investigated and considered, and provides future directions for our research.

We then work closely with domain experts in EHR data analysis to understand the needs of health data professionals. We design and develop empirical visual designs to address the identified needs, as well as tackle unsolved problems and challenges identified in the previous steps. The collaboration also enables us to have a deeper interpretation of EHR data sets, thus formulating a more suitable and user-centric visualisation solution. We also adopt an iterative development approach with expert feedback, to ensure that our research responds to the needs of clinicians and beyond. This thesis seeks to create visualisations that are not only theoretically sound but practically effective in overcoming the specific challenges posed by EHR data.

## 1.4 Contributions

This thesis makes significant contributions to the field of EHR Vis by addressing key challenges described in [Section 1.2](#) and introducing novel techniques that advance the state-of-the-art. The contributions span multiple areas, ensuring a comprehensive approach to tackling the unique demands of EHR Vis.

- **Diverse EHR Literature Sources:** A comprehensive and up-to-date survey [\[311\]](#) covering the field EHR Vis. The survey summarises and classifies 51 papers based on six reoccurring research themes and the Unified Medical Language System (UMLS), an attempt to bridge multiple disciplinary including visualisation, visual analytics, healthcare, biomedical science, and related disciplines.
- **EHR Data Acquisition:** A mini-survey [\[311\]](#) that curates 34 high-quality open access EHR data sources and data sets, serving as a valuable starting point for researchers entering the field.
- **EHR Data Scalability:** Novel techniques to address challenges in EHR data scalability. We employ state-of-the-art techniques to improve the performance of our visual designs and cope with the growth of EHR data. This is discussed in both [Chapter 4](#) and [Chapter 5](#).
- **EHR Data Diversity:** We address the challenge of handling heterogeneous EHR data by developing visual designs that can effectively visualise geospatial, unstructured text and long time series data. We discuss the visualisation of geospatial data in [Chapter 4](#), unstructured text data in [Chapter 3](#), long time series data in [Chapter 5](#) and simulation modelling data in [Chapter 6](#).
- **EHR Vis Evaluation:** The exemplary collaborations with EHR domain experts. We use an iterative process of design, development, and evaluation, which enables us to have a deeper understanding of the EHR data and the needs of healthcare professionals and deliver a more suitable visualisation solution. We combine qualitative and quantitative evaluation methods to assess the effectiveness of our work, as discussed in [Chapter 3](#), [Chapter 4](#), and [Chapter 5](#).

Collectively, these contributions showcase a holistic approach to EHR Vis, bridging gaps in existing methods and offering tools that address the specific needs of EHR analysis. By focusing on scalability, interactivity, and user relevance, the thesis provides

a foundation for future innovations in this critical field.

## 1.5 Thesis Structure

The rest of this thesis incorporates the following structure: [Chapter 2](#) presents a comprehensive survey of EHR Vis literature, including a mini-survey of 34 high-quality open-access EHR data sources and data sets. [Chapter 3](#) presents a novel letter-space visualisation tool, developed through an iterative collaboration with EHR domain experts, to support the exploration of the unstructured clinical text in a structured manner. We then present a novel hybrid cartogram layout algorithm in [Chapter 4](#), which incorporates topological elements into Demers cartograms. Through iterative collaboration with EHR domain experts, the resulting algorithm enhances the legibility, readability, and overall accuracy of EHR data visualised through Demers Cartogram.

[Chapter 5](#) describes a novel and scalable visual design, Time Series Map, which extracts events from long time series data and structures them into visual hierarchies. We collaborate with health data experts to evaluate the effectiveness of the design with two real-world data sets recorded by continuous glucose monitors. The evaluation results show that the Time Series Map can effectively support flexible exploration of long time series data, which is particularly beneficial for clinicians and researchers needing to identify trends, anomalies, and key events within long time series data sets.

In [Chapter 6](#), we present a special chapter that chronicles the design and development journey of EnsembleDashVis, a visualisation dashboard specifically crafted to support modellers in interpreting a simulation model utilised to forecast COVID-19 trends. The work took place amidst the exceptional circumstances of an unprecedented pandemic, and involved a fully remote and cross-disciplinary collaboration with over 40 domain experts from many fields such as data science, epidemiology, mathematics, public health, and many others.

In [Chapter 7](#), we conclude the thesis and provide future directions in EHR Vis. We provide all supplementary materials in [Section 7.2](#).



## Chapter 2

# EHR STAR: The State-Of-the-Art in Interactive EHR Vis

Wang, Q., & Laramée, R. S. (2022). EHR STAR: The State-Of-the-Art in Interactive EHR Vis. *Computer Graphics Forum*, 41(1), 69–105. <https://doi.org/10.1111/cgf.14424> [311]

*“A wealth of information creates a poverty of attention.”*

– Herbert A. Simon, Computer Scientist (1916 - 2001)

This chapter is based on the survey published in *Computer Graphics Forum* [311]. The content of this survey provides a solid foundation of visualisation knowledge as well as sources of EHR data for the rest of this thesis. Through the survey, we have identified the challenges and opportunities in the field of EHR Vis, and we have classified the existing EHR Vis techniques, with a mini survey on open access EHR data sets. A list of domain experts and potential collaborators was also identified through our intensive literature review.

## Contents

---

<b>2.1</b>	<b>Introduction and Motivation . . . . .</b>	<b>15</b>
2.1.1	Survey Challenges . . . . .	16
2.1.2	Literature Search Methodology . . . . .	17
2.1.3	Survey Scope . . . . .	22
2.1.4	Background and Terminology . . . . .	23
<b>2.2</b>	<b>Literature Classification . . . . .</b>	<b>27</b>
2.2.1	Multidisciplinary Research Themes . . . . .	27
2.2.2	Adopting a Medical Terminology Standard . . . . .	27
<b>2.3</b>	<b>Related Work . . . . .</b>	<b>29</b>
2.3.1	Related Work with an EHR Focus . . . . .	30
2.3.2	Related Work with a PopHR Focus . . . . .	35
<b>2.4</b>	<b>EHR Vis . . . . .</b>	<b>36</b>
2.4.1	Machine Learning . . . . .	37
2.4.2	Natural Language Processing . . . . .	40
2.4.3	Event Sequence Simplification . . . . .	42
2.4.4	Visual Analytics and Comparison . . . . .	45
2.4.5	Visual Analytics with Clustering and Others . . . . .	48
2.4.6	PopHR Vis and Geospatial Visualisation . . . . .	49
<b>2.5</b>	<b>Evaluation . . . . .</b>	<b>52</b>
<b>2.6</b>	<b>Open Access Healthcare Data . . . . .</b>	<b>55</b>
2.6.1	Healthcare Data Challenges . . . . .	55
2.6.2	Healthcare Data Search Methodology . . . . .	56
2.6.3	Healthcare Data Scope . . . . .	56
2.6.4	Healthcare Data Sources Classification . . . . .	57
2.6.5	Open Access Healthcare Data Sources . . . . .	58
<b>2.7</b>	<b>Future Research Challenges and Discussion . . . . .</b>	<b>63</b>
<b>2.8</b>	<b>Conclusions . . . . .</b>	<b>69</b>

---



## 2.1 Introduction and Motivation

We first germinated the idea of writing a state-of-the-art report (STAR) on EHR Vis in 2019, as the quantity of EHR Vis literature has grown since the last highly-cited survey published by Rind et al. in 2013 [126]. The landscape of EHR Vis has changed dramatically since then, with new techniques (both Vis and NonVis) being introduced and applied to solve EHR-related problems. The initial search showed that there was never a consensus on the definition of an EHR, let alone EHR Vis. A new survey paper would be a good opportunity for us to gain a comprehensive understanding of unsolved problems and provide the community with an up-to-date overview of the state-of-the-art in EHR Vis.

In this EHR STAR, we present literature reviews of papers related to EHR Vis from multiple disciplines, including visualisation, visual analytics, healthcare, and biomedical science. We attempt to define EHR Vis by combining the existing definitions and usage in 213 related references, and systematically extracting essential information from 51 papers based on multiple criteria. We also discuss the challenges and opportunities of EHR Vis, and provide a guide for future research. One major challenge is the access to EHR data sets, we specifically include a mini-survey of 34 high-quality open access EHR data sources in the STAR. Our contributions to the field 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 data sets.
- A STAR that appeals to researchers from visualisation, visual analytics, healthcare, biomedical science, and related disciplines.
- An overview of future challenges and open research directions in the field, for both new researchers and experts.

We have also developed an online EHR STAR literature browser for readers: <https://ehr.wangqiru.com>. It features all EHR papers and data sets along with several filtering and sorting options based on author, year, technique, and search terms. We believe that it offers a valuable

resource for those interested in this topic.

### 2.1.1 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 organise 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 visualisation, 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 2.2.2](#), and presenting a list of standardised terminology and definitions used in the related literature in [Section 2.1.4](#).

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 data sets for conducting research in the field, due to the sensitive nature of the data [\[175\]](#). There are a number of ways of acquiring electronic healthcare data sets:

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 data sets:** There are a number of open access data sets available online. In order to address this challenge directly, we classify and describe them in [Section 2.6](#). However, the challenge with such data sets is that 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 data sets are old and outdated.
3. **Proprietary data sets:** A license to access proprietary data sets can be expensive. We provide some example license costs in [Section 2.6.5](#) where we describe some proprietary data sets.

**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 [\[70\]](#) and GDPR [\[179\]](#). 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 [\[270\]](#).

**Scalability:** The size of an electronic healthcare data set is often huge. The rate of data growth exceeds the capacity of algorithms and software developed to visualise it [\[141\]](#).

**High-dimensionality:** Closely related to heterogeneity, healthcare data sets are high-dimensional and complex [\[132, 201\]](#). The ability to visualise large data sets with many attributes effectively remains a challenging problem [\[141\]](#).

We address some of these challenges directly in this STAR in [Section 2.6](#), which includes a survey of open access electronic healthcare data sources. We also present related future challenges in the field in [Section 2.7](#).

### 2.1.2 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 Transac-*

Source (Visualisation Venues)	Years	No. of Papers
IEEE Transactions on Visualisation 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 Visualisation 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 Visualisation 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
<b>Total</b>	<b>2006-2021</b>	<b>33</b>
Source (NonVisualisation 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
<b>Total</b>	<b>1998-2020</b>	<b>18</b>

Table 2.1: Conferences and Journals (both Visualisation and NonVisualisation venues) used for discovering literature and the number of papers found.

*tions on Visualisation 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 visualisation techniques for healthcare data

After the initial search and looking into the references, we found more literature from venues listed in [Table 2.1](#).

Search Keywords	Additional Keywords
Visualisation	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

Table 2.2: Keyword combinations used for discovering EHR Vis literature.

We first conduct a breadth-first search. Table 2.2 shows the list of keyword combinations we use for our breadth-first literature search. We use IEEE Xplore [341], The ACM Digital Library [347], Google Scholar [337], Vispubdata [196], Semantic Scholar [232], Mendeley [331] and Research Gate [359] 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 [198], we find another two related surveys on EHR Vis [126, 157].

UMLS Code	UMLS Term	Keywords	Number of Electronic Health Records Visualised					
			1	2 - 100	101 - 1,000	1,001 - 5,000	5,001 - 100,000	>100,000
C0003125	Anorexia nervosa			[39]				
C0004238	Atrial fibrillation			[148]				
C0007222	Cardiovascular diseases	Cardiovascular disease					[264]	
C0009378	Colonoscopy	Colonoscopy, biopsy, appendiceal-orifice			[219]			
C0010337	Care of intensive care unit patient	Critical care	[30]	[19] [248]		[136]	[231]	
C0011847	Diabetes		[79]		[91]			
C0011854	Diabetes mellitus, insulin-dependent	Type 1 diabetes	[251]		[265]			[238]
C0014544	Epilepsy				[298]			
C0018802	Congestive heart failure						[115]	

table continued on next page ...

...continued

UMLS Code	UMLS Term	Keywords	Number of Electronic Health Records Visualised					
			1	2 - 100	101 - 1,000	1,001 - 5,000	5,001 - 100,000	>100,000
C0020179	Huntington disease	Huntington's disease			[265]			
C0020443	Hypercholesterolemia							[238]
C0020538	Hypertensive disease	Hypertensive disease						[238]
C0021400	Influenza		[103] [125]					
C0021711	Neonatal intensive care		[134]	[19] [168]				
C0023981	Longitudinal Studies	Longitudinal cohort study				[222]		
C0024117	Chronic obstructive airway disease	Chronic obstructive pulmonary disease					[210]	
C0030567	Parkinson disease	Parkinson's disease			[265]			
C0030677	Patient care management		[33]			[136]		
C0030704	Patient transfer						[100]	
C0031330	Pharmacology	Pharmacovigilance			[122]			
C0031437	Phenotype			[163] [193]		[209]		
C0032285	Pneumonia		[103]					
C0034065	Pulmonary embolism		[30]					
C0035242	Respiratory tract diseases	Respiratory diseases					[264]	
C0038454	Cerebrovascular accident	Stroke			[171]			
C0040034	Thrombocytopenia				[71]			
C0085207	Gestational diabetes	Gestational diabetes mellitus		[148]				
C0262926	Medical history		[102]	[248]		[132]	[45]	[166] [238]
C0441472	Clinical action		[92]					

table continued on next page ...

...continued

UMLS Code	UMLS Term	Keywords	Number of Electronic Health Records Visualised					
			1	2 - 100	101 - 1,000	1,001 - 5,000	5,001 - 100,000	>100,000
C0599880	Treatment plan		[30] [89] [92] [251]				[264]	
C0600139	Prostate carcinoma	Prostate cancer				[224]	[144] [143]	
C0679831	Patient history	Patient's history	[11] [48] [189]				[144] [143]	
C0684249	Carcinoma of lung	Lung cancer	[128]					
C0872379	Disease subtype	Disease subtyping				[209]		
C1659543	Breast Density				[151]			
C2711227	Steatohepatitis	Hepatic Steatosis			[151]			
C3242284	Population health				[151]	[202] [201] [218] [241]		[113] [177]
C5204342	Clinical history	Patient clinical history	[228]	[148]				
UMLS Code	UMLS Term	Keywords	1	2 - 100	101 - 1,000	1,001 - 5,000	5,001 - 100,000	>100,000
Number of Electronic Health Records Visualised								

Table 2.3: **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 [32]. 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 visualised on the x-axis. Green highlights context papers included in this STAR.

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 [331] "related documents" functions.
- The "cited by" function provided by Google Scholar [337] and Semantic Scholar [232] to discover forward-looking related papers.

### 2.1.3 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.

#### In Scope

In this STAR, we focus on EHR and PopHR Vis as defined in [Section 2.1.4](#).

We include peer-reviewed literature focusing on real-world scenarios and empirical applications of EHR Vis. We emphasise 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 [\[122\]](#) hidden in EHR data-related processes. We include papers on EHR Vis with geospatial visualisation, as a geographical dimension might be relevant in a PopHR data set. Geospatial visualisation partially overlaps with this survey. We include research describing EHR Vis with Natural Language Processing (NLP) techniques. Friedman and Hripcsak recognise text visualisation with NLP as one of the most commonly used tools to extract information from EHR data and for studying clinical and research questions [\[14\]](#). 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 [\[182\]](#).

We focus on papers published in the previous 10 years. We refer to these papers as *focus papers*. Older papers such as LifeLines [\[8\]](#), LifeLines2 [\[11\]](#) and PatternFinder [\[45\]](#), contribute significantly to the field, with mature implementations deployed in clinical practises. We still include them as *context papers* and in the meta-data such as the classification [Table 2.3](#), 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. EventFlow [110] is a state-of-the-art system for visualising event sequences and exploring point and interval event patterns. Despite being influential in the field of EHR Vis, the work is excluded due to the absence of a peer-review process. However, we include a closely related paper by Monroe et al. [122] published in IEEE TVCG. We also exclude posters.

**Resource-oriented:** We exclude papers focusing on the visualisation of related resource-oriented EHR data. We define resource-oriented EHR data as the data that focuses on the management of clinical practises, such as hospital bed occupancy rates and inter/intra-hospital patient transfer times. These studies generally do not focus on clinical decision support directly. We exclude SepVis [186] as its focus is on the assessment of hospital performance based on the elapsed time between clinical activities and delays in clinical processes. We also exclude RadStream [204] as it is focusing on optimising the workflow in radiology departments. QualDash [258] is excluded due to its focus on adaptive dashboards for hospital quality improvement.

**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 visualisation 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 state-of-the-art.

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

## 2.1.4 Background and Terminology

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

Literature	EHR	EMR	Other Terms	Vis community	Year
Plaisant et al. [11]			Computerized patient records		1998
Horn et al. [19]				✓	2001
Bade et al. [30]					2004
Goren-Bar et al. [33]			Time-oriented clinical data	✓	2004
Hinum et al. [39]			Medical data		2005
Fails et al. [45]			Personal medical history	✓	2006
Bui et al. [48]				✓	2007
Wang et al. [71]				✓	2009
Rind et al. [79]			Medical data		2010
Faiola and Newlon [89]					2011
Gotz et al. [91]				✓	2011
Gschwandtner et al. [92]			Patient record	✓	2011
Wongsuphasawat et al. [100]				✓	2011
Zhang et al. [102]				✓	2011
Alonso and McCormick [103]			Public health data		2012
Sopan et al. [113]			Public health data		2012
Wongsuphasawat and Gotz [115]				✓	2012
Monroe et al. [122]				✓	2013
Ramírez-Ramírez et al. [125]			Public health	✓	2013
Borland et al. [128]			Population health	✓	2014
Gotz and Stavropoulos [132]				✓	2014
Kamaleswaran et al. [134]				✓	2014
Malik et al. [136]				✓	2014
Bernard et al. [144]				✓	2015
Bernard et al. [143]				✓	2015
Federico et al. [148]				✓	2015
Klemm et al. [151]			Population health	✓	2015
Glueck et al. [163]				✓	2016
Jiang et al. [166]				✓	2016
Kamaleswaran et al. [168]				✓	2016
Loorak et al. [171]				✓	2016
Ola and Sedig [177]					2016
Dabek et al. [189]				✓	2017
Glueck et al. [193]				✓	2017
Tong et al. [202]			Public healthcare data	✓	2017
Tong et al. [201]			Public healthcare data	✓	2017
Glueck et al. [209]				✓	2018
Guo et al. [210]				✓	2018

table continued on next page ...

... continued

Literature	EHR	EMR	Other Terms	Vis community	Year
Tong et al. [218]			Public healthcare data	✓	2018
Trivedi et al. [219]					2018
Alemzadeh et al. [222]			Longitudinal cohort study	✓	2019
Bernard et al. [224]				✓	2019
Glicksberg et al. [228]					2019
Guo et al. [231]				✓	2019
Kwon et al. [238]				✓	2019
McNabb and Laramee [241]			Population health data	✓	2019
Sultanum et al. [248]				✓	2019
Zhang et al. [251]				✓	2019
Jin et al. [264]					2020
Kwon et al. [265]				✓	2020
Wang et al. [298]				✓	2021
<b>Total unique papers: 51</b>	25	10	16	40	

Table 2.4: **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 2.3. This table indicates a mixture of terms is used throughout the literature. We clarify the terminology in Section 2.1.4. Green highlights context papers.

**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 [175]. Iakovidis defines EHR as digitised healthcare information on individual patients that is accessible, secure and highly usable for supporting the analysis of healthcare, education and research [10]. Gunter and Terry define EHR as, “A longitudinal collection of electronic health information about individual patients and populations” [38, 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” [342]. 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 person’s care under a particular provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunisations, laboratory data and radiology reports” [349]. The World Health Organisation (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*” [199, 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 digitised health records used to improve quality of care and estimate costs [102, 161, 192, 217]. Unlike EHR, an EMR is stored and used internally without inter-organisation sharing [195]. 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. [47] stating that a PHR differs from an EHR by its accessibility. A PHR is managed by the data owner and is authorised for sharing with healthcare providers when necessary [47]. 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*” [343]. 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 [330].

**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 authorised users*” in 2010 [74, p.360]. A PopHR data set 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 summarised in [Section 2.4.6](#).

**EHR Vis:** We consider the visualisation of EHR and PopHR for clinical decision support, as a sub-field of information visualisation and visual analytics (EHR Vis).

## 2.2 Literature Classification

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

- Recurring multidisciplinary research themes derived from our literature search, described in [Section 2.2.1](#).
- The Unified Medical Language System (UMLS), introduced in [Section 2.2.2](#), as the medical terminology standard for classifying literature.

### 2.2.1 Multidisciplinary Research Themes

EHRs are often large-scale and may contain noisy data [206]. This means an automated process can be implemented in order to achieve both efficiency and accuracy in the preprocessing and visualisation stages. From the related literature, we have identified several major research themes in processing and visualising EHRs. We provide a brief description of these themes here and review the related literature in detail in [Section 2.4](#).

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

[Table 2.8](#) shows an overview of our literature classification based on multidisciplinary research themes.

### 2.2.2 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 visualisation [192]. By adopting a medical terminology standard, we hope to bridge the gap between two communities, thus reaching a wider audience beyond information visualisation and

visual analytics, and taking advantage of the extensive work invested into standardised terminology development.

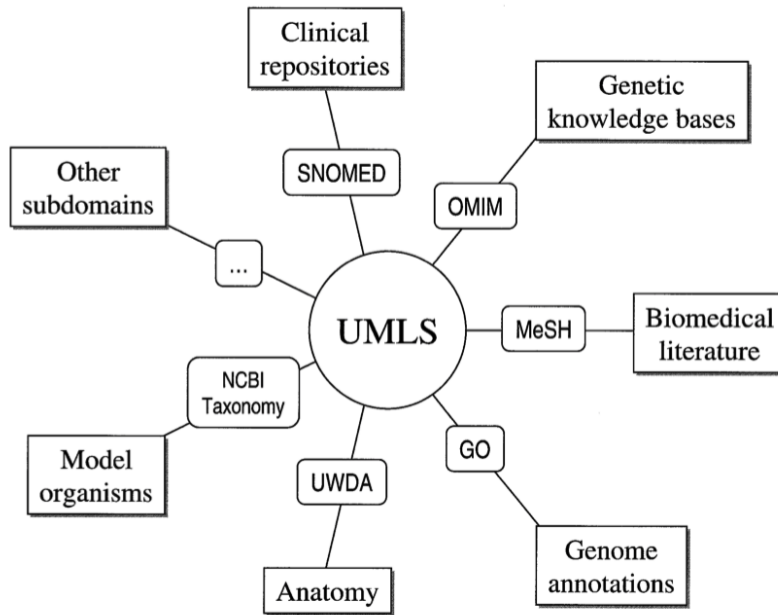


Figure 2.1: The various subdomains integrated in the UMLS Terminology. Figure reproduced from Abel [32].

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 standardisation. A schematic of the integrated dictionaries is shown in Figure 2.1. Dictionaries often use different lexical items to describe identical or similar terms. An integrated standard will make these resources interoperable, and machine-readable and help dismantle the barrier to multidisciplinary research [32].

In order to classify each paper, we first extract their keywords to obtain their corresponding code and terminology from the UMLS. Table 2.3 shows the overview classification of research papers found in our literature search. The x-axis is mapped to the number of EHRs visualised 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 2.3 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 in a wider team effort to tackle challenges in the field. And

Literature	Related Work							Year
	Ⓔ Rind et al. [126]	Ⓟ Carroll et al. [129]	Ⓔ West et al. [157]	Ⓔ Gotz and Borland [164]	Ⓔ Onukwugha et al. [178]	Ⓔ Rind et al. [200]	Ⓟ Preim and Lawonn [270]	
Gotz et al.[91]								2011
Wongsuphasawat et al.[100]								2011
Alonso and McCormick[103]								2012
Sopan et al.[113]								2012
Wongsuphasawat and Gotz[115]								2012
Monroe et al.[122]								2013
Ramírez-Ramírez et al.[125]								2013
Borland et al.[128]								2014
Malik et al.[136]								2014
Bernard et al.[144]								2015
Bernard et al.[143]								2015
Federico et al.[148]								2015
	4	3	4	2	3	3	2	
Total unique papers: 12   Total appearances: 21								

Table 2.5: **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.

finally, we can observe that not many papers are dealing with the really large EHR and PopHR data sets with over 100,000 records.

## 2.3 Related Work

This section introduces related work with a special emphasis on previous related surveys. Papers with a focus on visualisation or visual analytics of EHR data are described in [Section 2.3.1](#). We present previous PopHR survey papers in [Section 2.3.2](#).

Our STAR differs from previous ones by including a novel, up-to-date overview using a medical terminology standard described in [Section 2.2.2](#), with 29 more recent publications on EHR Vis. [Table 2.5 - 2.7](#) 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 2.6](#) to address the challenge of healthcare data access.

Literature	Related Work						Year
	Roque et al. [80]	Rind et al. [126]	Simpao et al. [140]	West et al. [157]	Onukwugha et al. [178]	Rind et al. [200]	
Plaisant et al.[11]							1998
Horn et al.[19]							2001
Bade et al.[30]							2004
Goren-Bar et al.[33]							2004
Hinum et al.[39]							2005
Fails et al.[45]							2006
Bui et al.[48]							2007
Wang et al.[71]							2009
Rind et al.[79]							2010
Faiola and Newlon[89]							2011
	4	9	1	3	1	2	
Total unique papers: 10   Total appearances: 20							

Table 2.6: **Context papers:** Y-axis, overlapping context papers from previous survey papers, ordered by the year of publication. X-axis,  $\textcircled{E}$  indicates an EHR focused survey and  $\textcircled{P}$  indicates a PopHR focused survey. We can observe that the 2013 survey by Rind et al. [126] has some thematic overlap with this one.

### 2.3.1 Related Work with an EHR Focus

In this section, we divide related work with an EHR focus into two subcategories, 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 2.5 - 2.7 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 [161]. The quantity of literature has grown since an early survey published in 2013 by Rind et al. [126]. There are a number of older related surveys published since then, we present them in this section.

Roque et al. compare six information visualisation systems designed for providing













Literature	Related Work								Exclusion Criteria	Year
	 Roque et al. [80]	 Rind et al. [126]	 Carroll et al. [129]	 Simpao et al. [140]	 West et al. [157]	 Onukwugha et al. [178]	 Rind et al. [200]	 Preim and Lawonn [270]		
Kosara and Miksch[21]									B	2001
Atkinson and Unwin[24]									OS	2002
Chittaro et al.[27]									OT	2003
Brodbeck et al.[37]									B	2005
Aigner and Miksch[42]									B, OS	2006
Blanton et al.[43]									OT	2006
Da Silva et al.[49]									B, OS	2007
Guo[51]									B, OT	2007
Hu et al.[52]									B, OT	2007
Pieczkiewicz et al.[53]									B	2007
Gao et al.[54]									B, OS	2008
Hallett[55]									B	2008
Heitgerd et al.[56]									B	2008
Reinhardt et al.[59]									B	2008
Yi et al.[63]									OS	2008
Bashyam et al.[64]									B, S	2009
Connors et al.[65]									OT	2009
Wongsuphasawat and Shneiderman[72]									OT	2009
Goldsmith et al.[75]									B	2010
Klimov et al.[76]									RO	2010
Kumasaka et al.[77]									OT	2010
Naumova[78]									OT	2010
Steenwijk et al.[81]									S	2010
Willison[85]									B, N, OS	2010
Driscoll et al.[88]									B, OT	2011
Hripsak et al.[93]									B, OS	2011
Lewis et al.[96]									B	2011
Maciejewski et al.[97]									B, OT	2011
Gesteland et al.[104]									OT	2012
Joshi and Szolovits[107]									B, OS	2012
Livnat et al.[108]									B, OT	2012
Mane et al.[109]									B	2012
Perer and Sun[111]									B, OS	2012
Stubbs et al.[114]									B	2012
Rajwan et al.[124]									B, OS	2013
Freifeld et al.[130]									B, OS	2014
Gálvez et al.[131]									B	2014
Simpao et al.[140]									B	2014
Dunne et al.[147]									B, OT	2015
Masoodian et al.[172]									B, OT	2016
Caballero et al.[186]									RO	2017
Abukhodair et al.[204]									RO	2018
<div>3101178416</div> <div>Total unique papers: 42   Total appearances: 50</div>										

Table 2.7: **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 2.1.3: (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.

overviews of EHR data [80]. Systems are classified based on the users, goals, and tasks. Four of these systems are included in our survey as context papers (Table 2.6) and two are excluded for reasons indicated in Table 2.7.

Rind et al. [126] review 14 information visualisation 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 visualisation has in supporting medical tasks. Some 14 systems are compared by (1) supported data types (categorical and numerical), (2) multivariate 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 2.5), nine are included as context papers (Table 2.6), and seven are papers considered out of scope with reasoning indicated in Table 2.7.

Simpao et al. [140] discuss applications of visual analytics in healthcare since the HITECH Act in 2009. The authors review eight visual analytics tools for EHR and categorise their application 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, optimising resource allocation. From these eight EHR Vis tools, one is included in our survey as a context paper (Table 2.6), and seven are considered out of scope (Table 2.7).

West et al. [157] publish a systematic review of 18 papers, by highlighting crucial metrics to evaluate EHR systems. Those metrics include (1) visualisation techniques applied to utilise the screen space efficiently while preserving as much data as possible, (2) interactive user options to identify abnormalities within the data, (3) visualisation of the entire data set even if there are missing values or inaccurate data entries, (4) visualisation 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 2.5), three as

context papers (Table 2.6) and exclude six papers (Table 2.7).

Onukwugha et al. [178] 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 visualising a prostate cancer archive and a health insurance claim data set. The authors focus on EHR systems from three perspectives, (1) the ability to identify and rectify errors in data, (2) visualisation techniques and interactive options provided to support data analysis, and (3) cogent visualisations generated to present findings to decision makers. From these seven EHR Vis systems, we include four as focus papers (Table 2.5), one as a context paper (Table 2.6) and exclude two papers (Table 2.7).

Gotz and Borland [164] discuss challenges and opportunities for the interactive visualisation 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, provide 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 supports discovery and insight that generalise across a population at large. We include two as focus papers (Table 2.5) in our survey and exclude two papers (Table 2.7).

Rind et al. [200] publish a survey of EHR Vis with a focus on time-oriented data sets. The authors identify technical challenges arising from the temporal dimension of EHR data sets, 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 2.5) and two as context papers (Table 2.6).

## Related Work with EHR Focus Outside the Visualisation 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 visualisation

community with a focus on EHR data.

MIT Critical Data published a related book, *Secondary Analysis of Electronic Health Records* [175]. The first chapter identifies the objective of secondary analysis of EHR data as the utilisation of EHR data to provide evidence to inform best practice 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 2.17 in Section 2.6.5 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 data sets 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 2.7 and Table 2.18. Other identified challenges are data protection, data interoperability, the cost of data infrastructure, and the varied quality of research output. The rest of the book describes techniques in data preprocessing and analysis with example studies conducted using EHR databases reviewed in chapter two.

Shickel et al. [217] 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. [217] survey. Another classification dimension is derived from the target task and subtasks of previous EHR systems.

Koleck et al. [236] systematically review 27 systems that adopt NLP algorithms for extracting structured data from free text EHRs. Table 3 in Koleck et al. [236] shows the classification by clinical speciality. 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 improves the effectiveness of tailored interventions. Some 27 systems are evaluated, with eight critical indicators identified by the authors.

### 2.3.2 Related Work with a PopHR Focus

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

Carroll et al. [129] publish a systematic review of 88 articles with a primary focus on infectious disease, needs of public health users, or usability of information visualisations. 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 2.5) as focus papers, none as context papers, and exclude 11 with reasons indicated in Table 2.7.

Preim and Lawonn review the existing visual analytics solutions for supporting Public Health (PH) [270] with structured data. The authors describe PH data sets 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 visualisation 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 2.5) as focus papers, none as context papers, and exclude six with reasons indicated in Table 2.7.

## 2.4 EHR Vis

This section describes 41 focus papers on EHR Vis found from our literature search. We further categorise these papers based on six multidisciplinary research themes derived from our investigation, as shown in Table 2.8. 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 guidelines provided by Laramee ([95]).

Literature	Complementary Techniques							Year
	ML	NLP	ESS	GEO	Clustering	Comparison	Others	
Plaisant et al. [11]								1998
Horn et al. [19]								2001
Bade et al. [30]								2004
Goren-Bar et al. [33]								2004
Hinum et al. [39]								2005
Fails et al. [45]								2006
Bui et al. [48]								2007
Wang et al. [71]								2009
Rind et al. [79]								2010
Faiola and Newlon [89]								2011
Gotz et al. [91]								2011
Gschwandtner et al. [92]								2011
Wongsuphasawat et al. [100]								2011
Zhang et al. [102]								2011
Alonso and McCormick [103]								2012
Sopan et al. [113]								2012
Wongsuphasawat and Gotz [115]								2012
Monroe et al. [122]								2013
Ramírez-Ramírez et al. [125]								2013
Borland et al. [128]								2014
Gotz and Stavropoulos [132]								2014
Kamaleswaran et al. [134]								2014
Malik et al. [136]								2014
Bernard et al. [144]								2015
Bernard et al. [143]								2015
Federico et al. [148]								2015
Klemm et al. [151]								2015

table continued on next page ...

... continued

Literature	Complementary Techniques							Year
	ML	NLP	ESS	GEO	Clustering	Comparison	Others	
Glueck et al. [163]						Red		2016
Jiang et al. [166]		Grey		Red				2016
Kamaleswaran et al. [168]					Red			2016
Loorak et al. [171]			Red			Grey		2016
Ola and Sedig [177]				Red				2016
Dabek et al. [189]	Red							2017
Glueck et al. [193]		Grey				Red		2017
Tong et al. [202]				Red				2017
Tong et al. [201]				Red				2017
Glueck et al. [209]	Red					Grey		2018
Guo et al. [210]	Grey	Grey	Red					2018
Tong et al. [218]				Red				2018
Trivedi et al. [219]	Grey	Red						2018
Alemzadeh et al. [222]				Red				2019
Bernard et al. [224]			Red					2019
Glicksberg et al. [228]							Red	2019
Guo et al. [231]	Grey		Red					2019
Kwon et al. [238]	Red							2019
McNabb and Laramee [241]				Red				2019
Sultanum et al. [248]	Grey	Red						2019
Zhang et al. [251]						Red		2019
Jin et al. [264]	Red		Grey					2020
Kwon et al. [265]	Red							2020
Wang et al. [298]		Grey				Red		2021
<b>Total unique paper: 51</b>	10	8	12	11	3	10	8	

Table 2.8: **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, **Grey** highlights the secondary theme, and **Green** highlights context papers.

### 2.4.1 Machine Learning

This section introduces the literature that combines Machine Learning (ML) and EHR Vis. We follow the definition of ML by Alpaydin [73] as the process of optimising the performance of a predefined model, based on example data or past experience.

Literature	ML Topics	UMLS Term	Year
Bernard et al.[144]	Active LearningREPTree	Prostate carcinoma	2015
Dabek et al.[189]	Unspecified	Patient history	2017
Guo et al.[210]	Clustering	Chronic obstructive airway disease	2019
Glueck et al.[209]	Active Learning	Phenotype Disease subtype	2018
Trivedi et al.[219]	Support Vector MachineBag-of-words	Colonoscopy	2018
Guo et al.[231]	Neural Networks	Care of intensive care unit patient	2019
Kwon et al.[238]	Recurrent Neural Networks	Diabetes mellitus, insulin-dependent Hypercholesterolemia Hypertensive disease Medical history	2019
Sultanum et al.[248]	Active Learning	Care of intensive care unit patient Medical history	2019
Jin et al.[264]	Recurrent Neural Networks	Cardiovascular diseases Respiratory tract diseases Treatment plan	2020
Kwon et al.[265]	Hidden Markov Models	Diabetes mellitus, insulin-dependent Huntington disease Parkinson disease	2020

Table 2.9: An overview table of ML topics discussed in the literature described in Section 2.4.1. Papers with ML as a secondary theme are highlighted in Green.

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 [215], neural networks [238], support vector machines [250, 225] and topic models [225], have evolved recently to increase automation of processing EHR archives. From examining Table 2.8, 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 2.9 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 visualisation literature.

Bernard et al. contribute a visual active learning system [143] extending their prior work [144]. The system enables 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 visualisation enables quick identification of possible diagnoses of individual patient’s symptoms.

Dabek et al. propose a timeline-based framework for aggregating and summarising



EHRs [189]. 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 visualising patient diagnoses and a horizon chart visualising lab test results.

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 [209]. Based on the Human Phenotype Ontology (HPO) extracting and mapping method introduced in the prior work [163, 193], PhenoLines aims to support the filtering, comparison, simplification and interpretation of temporal evolution of phenotype probabilities within and between disease subtypes. Topic modelling 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 2.2.



Figure 2.2: PhenoLines [209] 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. Figure reproduced from Glueck et al. Glueck et al. [209].

Kwon et al. [238] 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

Literature	NLP Approaches	UMLS Term	Year
Zhang et al.[102]	Unspecified	Medical history	2011
Jiang et al.[166]	Named Entity Recognition	Medical history	2016
Glueck et al.[193]	Natural Language Queries	Phenotype	2017
Trivedi et al.[219]	Automated Retrieval Console cTAKES	Colonoscopy	2018
Sultanum et al.[248]	cTAKES	Care of intensive care unit patient Medical history	2019
Wang et al.[166]	Natural Language Queries	Epilepsy	2021

Table 2.10: An overview table of NLP approaches adopted by the literature described in Section 2.4.2. Papers with NLP as a secondary theme are highlighted in Green.

prediction.

Jin et al. [264] 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 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 data set [167] is used for thorough evaluations with seven physicians including two case studies. We include this open access data set in Table 2.17.

Kwon et al. [265] present DPVis, a multiple views visual analytics system that focuses on visual disease progression analysis in order to develop fully interpretable and interactive visualisations. 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.

## 2.4.2 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 analysing 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 applica-*

tions” [29]. 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 [236]. Clinicians have very diverse ways of documenting patient records. This may require appropriate modifiers to capture words, phrases and their relationships in EHRs [248]. Table 2.10 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.

Zhang et al. develop AnamneVis in order to capture a complete picture of a patient’s medical history [102]. 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) is the medical standard for mapping diseases and symptoms. The Five Ws concept [127] is adopted for mapping the relations between extracted information. A sunburst diagram is used to visualise the data in two layouts, (1) a hierarchy-centric layout for the 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 the past medical diagnostic flow of the patient.

Trivedi et al. introduce NLPReViz, a visual analytic and visualisation tool that uses Support Vector Machine for training NLP models in real-time [219]. 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 the 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 2.8.

Sultanum et al. present Doccurate, a system embodying a curation-based approach that automatically extracts relevant information from large clinical text data sets, to provide an accurate and sufficient overview for a patient [248]. After interviewing six domain experts, the authors conclude that preserving the original text in clinical notes is crucial for the visualisation of EHRs. Doccurate provides automation in data processing and customisation for visualisation while preserving the link to the original

data.

### 2.4.3 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 [123, 173]. EHRs by nature are temporal events unfolding successively, ESS enables events to be trimmed down to their core elements, improving both data-processing and visualisation of EHRs. The technique is adopted by EHR Vis systems such as LifeLines [8] and EventFlow [122]. Table 2.11 provides a summary of event types appearing in this sub-section. Events associated with hospitals are a recurring theme.

Wongsuphasawat et al. [100] introduce LifeFlow for providing an interactive visual overview of event sequence data. Following the approach used in LifeLines2 [71], 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 colour-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.

Wongsuphasawat et al. [115] introduce Outflow, a visualisation of temporal event sequence data. Outflow uses a different approach that visualises 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.

Monroe et al. introduce a technique to simplify temporal event sequence data [122], following their previous work called EventFlow [110]. 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 2.3. 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 aggregates 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.

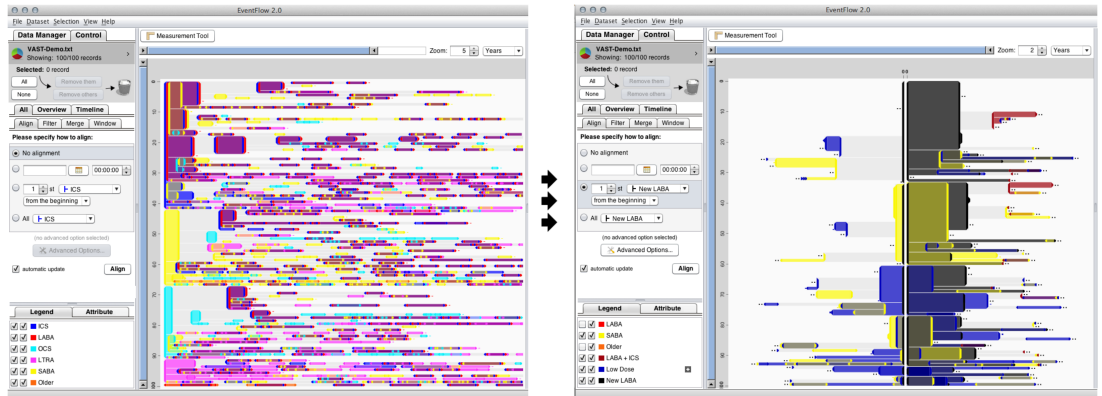


Figure 2.3: EventFlow [122] visualising the original Long-Acting  $\beta$ -Agonists data set on the left, and the simplified data set on the right. The number of visual elements is reduced from 2,700 to 492. Figure reproduced from Monroe et al. [122].

Gotz and Stavropoulos [132] introduce DecisionFlow for visualising large numbers (thousands) of high-dimensional temporal event sequence data. Instead of visualising the entire data set 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 visualised for further analysis and interactions. The user is able to set and modify milestones interactively to achieve filtering and selection.

Malik et al. present CoCo [136], a visual analytics tool for comparing groups (cohorts) of temporal event sequence data. Inspired by EventFlow [122] and Outflow [115], CoCo enables users to explore statistics about the underlying data set 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 [152] with two case studies.

Literature	Event Types	UMLS Term	Year
Wongsuphasawat et al.[100]	Hospital discharge and transfer flows	Patient transfer	2011
Wongsuphasawat and Gotz[115]	Congestive heart failure	Congestive heart failure	2012
Monroe et al.[122]	Prescriptions	Pharmacology	2013
Gotz and Stavropoulos[132]	Diagnoses, lab tests, and medications	Medical history	2014
Malik et al.[152]	Respiratory and radiation		2015
Loorak et al.[171]	Stroke	Cerebrovascular accident	2016
Guo et al.[210]	Diagnoses, procedures, hospital admission and discharge	Chronic obstructive airway disease	2018
Bernard et al.[224]	Biological indicator for prostate cancer	Prostate carcinoma	2019
Guo et al.[231]	Hospital admission and discharge, death, prescriptions, infusions, lab tests	Care of intensive care unit patient	2019
Jin et al.[264]	Hospital admission, prescriptions, diagnoses, treatments	Cardiovascular diseases Respiratory tract diseases Treatment plan	2020

Table 2.11: An overview table of event types in the literature described in Section 2.4.3. Papers with ESS as a secondary theme are highlighted in Green.

Loorak et al. [171] present TimeSpan, a visualisation 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 domain of stroke care analysis. Temporal events are visualised using a parallel coordinates with stacked bar charts extended with the Bertin-style matrices [12], and are aligned based on their positive effect on the patient. A unique evaluation with a focus group session is also presented.

Guo et al. describe EventThread [210], a visualisation system for revealing the evolution of patterns across stages in event sequence data. EventThread uses Term Frequency - Inverse Document Frequency (TF-IDF) [4], 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 facilitate further analysis. Guo et al. introduce EventThread 2 [231] 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.

Bernard et al. propose a technique for visualising post-operative prostate cancer, that segments patient histories based on time and then aggregates the results by therapy states and biological conditions [224]. Instead of treating patient histories as event

sequences, the segmented results are presented using a static dashboard, with extensive use of colours and glyphs for encoding variables, in order to visualise longitudinal changes in patient histories. The segmentation of patient histories is done by using a sliding window approach to traverse the data set. Evaluation is performed with groups of both expert and nonexpert users.

#### 2.4.4 Visual Analytics and Comparison

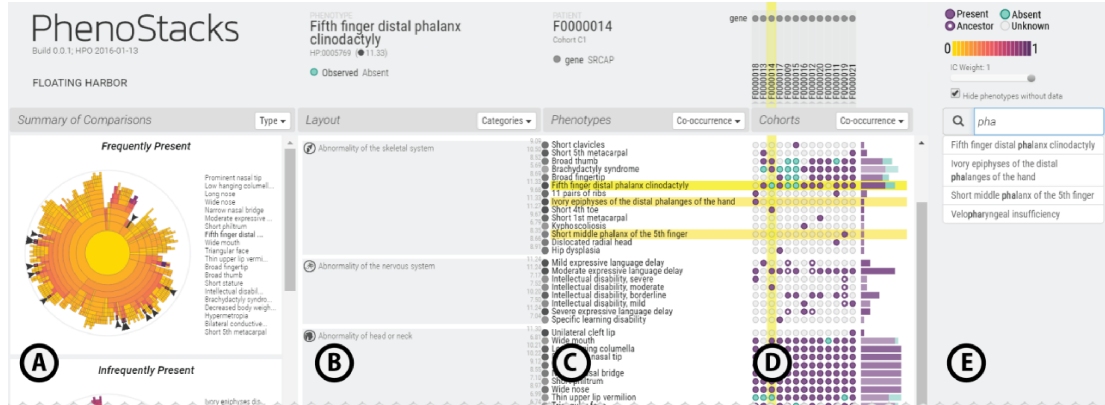


Figure 2.4: PhenoStacks [193] includes (A) The summary panel conveying phenotype patterns across patient cohorts in a sunburst chart, (B) The layout view enables the user to select phenotypes by collapsing, filtering, and clustering, (C) The list view shows the phenotype names with a sorting function, (D) The observations Plot visualises 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. Figure reproduced from Glueck et al. [193].

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. [90], juxtaposition, superposition, and explicit encodings. Table 2.12 summarises the types of comparison techniques used in the EHR Vis literature. Juxtaposition, the simplest, is the most common choice by a wide margin.

Gschwandtner et al. present CareCruiser [92], an enhanced visual analysis system to explore the result of each applied clinical action and identifies sub-optimal treatment choices. CareCruiser supports the visualisation 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

Literature	Comparative Design			UMLS Term	Year
	Juxtaposition	Superposition	Explicit Encoding		
Gschwandtner et al.[92]				Clinical action Treatment plan	2011
Borland et al.[128]				Carcinoma of lung	2015
Bernard et al.[143]				Prostate carcinoma Patient History	2015
Federico et al.[148]				Atrial fibrillation Gestational diabetes Clinical history	2015
Glueck et al.[163]				Phenotype	2016
Loorak et al.[171]				Cerebrovascular accident	2016
Glueck et al.[193]				Phenotype	2017
Glueck et al.[209]				Phenotype	2018
Glicksberg et al.[228]				Clinical history	2019
Zhang et al.[251]				Diabetes mellitus, insulin-dependent Treatment plan	2019
Wang et al.[298]				Epilepsy	2021

Table 2.12: An overview table of comparative designs adopted by the literature described in Section 2.4.4. The x-axis is mapped to the comparative design categorisation by Gleicher et al. [90]. Papers with Comparison as a secondary theme are highlighted in Green.

the patient’s condition and responses to treatments, as well as the comparison between multiple patients.

Borland et al. [128] describe radial coordinates, a visualisation technique based on parallel coordinates, a scatterplot, and a chord diagram. The technique allows more efficient utilisation 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 the comparison of high and low prevalence values across all dimensions in the data. The radial style parallel coordinates visual design is applied to NHS data from the UK.

Bernard et al. present an interactive visualisation system for identifying, categorising, and analysing EHRs of cohorts of prostate cancer patients [143]. The system supports the visualisation 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.

Federico et al. introduce Gnaeus, a guideline-based knowledge-assisted visual ana-



lytics system for EHRs [148]. Gnaeus utilises 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 visualise the outcome, with related CIGs on the side to provide recommendations. The system enables the user to compare administered treatment with evidence-based best practises.

Glueck et al. introduce PhenoBlocks, a visual analytics tool that supports the comparison of phenotypes between patients [163]. PhenoBlocks introduces a differential hierarchy comparison algorithm for analysing phenotypes pairwise between patients and uses a customised sunburst radial hierarchy layout [18] for visualising the results.

Glueck et al. present PhenoStacks, a visualisation system to support the comparison of cross-sectional phenotype within and between patient cohorts [193]. The system adopts glyphs of Human Phenotype Ontology (HPO) developed in the prior work PhenoBlocks [163] and supports sorting and filtering by phenotype or patient attributes. Search is powered with natural language queries (See Figure 2.4). To reduce visual redundancy, the authors propose a topology simplification algorithm, a greedy depth-first approach, for eliminating duplicates in phenotype data sets.

Zhang et al. describe IDMVis [251], a temporal event sequence visualisation 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 visualising temporal event sequences [190], the authors propose a visual technique of dual sentinel event alignment and time scaling to further enhance the visualisation 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.

Wang et al. present LetterVis, a visualisation tool to support the analysis of clinic letters through five customised visual layouts with support from natural language queries [298]. A letter-space layout is derived from the physical layout of text on A4-size letters used by clinicians, exploiting the implicit knowledge of the clinicians who compose the letters. This layout is used to depict the query results in (1) the

Literature	Clustering Dimensions	UMLS Term	Year
Gotz et al.[91]	Medical decisions	Diabetes	2011
Kamaleswaran et al.[134]	Temporal	Neonatal intensive care	2014
Kamaleswaran et al.[168]	Temporal Respiratory physiologic signals	Neonatal intensive care	2016

Table 2.13: An overview table of clustering dimensions used in the literature described in [Section 2.4.5](#).

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 included for visualising 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.

## 2.4.5 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 [41], hierarchical clustering algorithms are widely used in the information visualisation discipline. This conforms with our findings that all papers included in this section ([Table 2.13](#)) adopt hierarchical clustering algorithms to produce homogeneous subgroups based on similarities. EHRVis may benefit from applying other clustering techniques (e.g. Vector Quantisation and Estimation via Mixture Densities) to assist in analysis.

Gotz et al. introduce DICON [91], a visualisation 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.

Kamaleswaran et al. use a tri-event heatmap representation for displaying high-frequency complex data [134], neonatal spells, collected in neonatal intensive care units. Their clustering includes a temporal factor and a nonlinear similarity metric. The authors apply both density estimation and logarithmic clustering to normalize and discretise the nonparametric distribution during data preprocessing. The resulting

visualisation supports the exploration of the frequency, duration, and severity of spells.

Kamaleswaran et al. introduce a visualisation technique called a Temporal Intensity Map (TIM) [168], a customised 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 information about the 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 the physiological behaviours of newborns.

Glicksberg et al. describe PatientExploreR [228], an interactive interface that facilitates the visualisation and querying of EHRs. By incorporating the Observational Medical Outcomes Partnership (OMOP) common data model introduced by the Observational Health Data Sciences and Informatics [274], 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 2.2.2. 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.

## 2.4.6 PopHR Vis and Geospatial Visualisation

This section describes research on EHR Vis with a geospatial focus. Table 2.14 summarises the geospatial landscape coupled by this sub-section of literature. PopHR Vis papers are also included in this section. Alonso and McCormick describe Epidemiological Parameter Investigation from Population Observations Interface (EPIPOI), which automatically extracts three parameters describing trends, seasonality and anomalies, and a time series from large epidemiological data sets [103]. These three dimensions can be visualised using maps combined with time series data to reveal spatial patterns. 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.

Literature	Geospatial Regions	UMLS Term	Year
Alonso and McCormick[103]	Brazil	PneumoniaInfluenza	2012
Sopan et al.[113]	US	Population health	2012
Ramírez-Ramírez et al.[125]	Ontario, Canada	Influenza	2013
Klemm et al.[151]	Germany	Population health Breast Density Steatohepatitis	2013
Jiang et al.[166]	Indiana, US	Medical history	2016
Ola and Sedig[177]	Global	Population health	2016
Tong et al.[202]	England, UK	Population health	2017
Tong et al.[201]	England, UK	Population health	2017
Tong et al.[218]	England, UK	Population health	2018
Alemzadeh et al.[222]	Germany	Longitudinal studies	2019
McNabb and Laramée[241]	Ireland UK US	Population health	2019

Table 2.14: An overview table of geospatial regions covered in the literature described in [Section 2.4.6](#).

Sopan et al. introduce the Community Health Map [113] that interactively visualises public healthcare data sets using a multivariate choropleth map. Selection enables users to visualise multiple data sets 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.

Ramírez-Ramírez et al. introduce SIMID [125], a surveillance and spatio-temporal visualisation tool for infectious diseases. Based on the existing data, SIMID simulates the spread of infectious diseases using interactive animated maps. With customisable 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.

Jiang et al. introduce Health-Terrain [166] to support the visual exploration of large healthcare data sets. Based on UMLS described in [Section 2.2.2](#), 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, which consists of (1) constructing random noise patterns with colour variations to map different attributes, and (2) colour-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 [17] and spiral pattern [23], to help physicians discover patterns and trends in events. Health-Terrain is included in this section since it is a combination of geospatial visualisation and NLP with the main focus on the former.

Klemm et al. [151] propose the 3D Regression Heat Map, a novel 3D visual encoding that offers an overview of a hepatic steatosis data set (a subset of the SHIP data set included in Table 2.17). 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.

Ola and Sedig [177] present a geospatial visual design for studying large healthcare data sets. The design combines several visualisation techniques to support the exploration of the relationships between age group, risk, and cause of death at multiple levels of granularity.

Tong et al. present a hybrid visual layout called a cartographic treemap, to visualise high-dimensional healthcare data collected by the National Healthcare Service (NHS) in the U.K. [202]. 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 by adding a time variate, enabling the visualisation of the temporal evolution trends hidden in EHR data [201].

Tong et al. extend their previous work with a cartographic layout algorithm that generates cartograms with topological features using NHS’s population healthcare data [218]. The proposed algorithm preserves the nearby node’s topological features to increase the recognisability and reduce layout errors.

VIVID is a web-based framework proposed by Alemzadeh et al. [222] 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 bar chart and matrix) build imputation models (bean plot and bee swarm plot) and generate predictions for the missing values (chord diagram and parallel coordinates).

McNabb and Laramée present a glyph placement algorithm to support multivariate geospatial visualisation of a Public Health England data set [241]. 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) Visualisation 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 that are dynamic, scale-aware, and coupled to their geospatial contexts.

## 2.5 Evaluation

Evaluation of EHR and PopHR visual designs is very difficult due to their complex visual interfaces. An EHR Vis system is often characterised according to target user requirements. The resulting visual designs may not seem useful to evaluators [68]. Furthermore, an isolated evaluative process is hardly sufficient to assess an EHR Vis system. *Grounded evaluation* [57], where visualisation designers work closely with EHR experts to 1) understand pre-design 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 practised in many projects (especially from the visualisation community) included in this STAR.

In this section, we summarise the evaluation techniques adopted by each paper. The result is summarised in Table 2.15.

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 [121]. 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 ex-

perts participate in multiple stages of the software development life cycle, such as planning, requirement analysis, and testing [219, 224, 248, 264]. 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 the adoption of interviews since 2014. Interviewees usually respond in depth during interviewing sessions [163, 169, 193, 219, 224, 264] and provide personal interpretations beyond interaction and usability aspects [165].
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 [238]. 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's side or the reviewer's 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 [45, 72, 91, 136, 224]. 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 [121]. Controlled user studies are difficult to design for complex systems.

Literature	Evaluation				Year
	Domain Expert Feedback	Interview	Case Study	Controlled User Study	
Plaisant et al. [11]					1998
Horn et al. [19]	2				2001
Bade et al. [30]					2004
Goren-Bar et al. [33]	●				2004
Hinum et al. [39]	●	●			2005
Fails et al. [45]	8				2006
Bui et al. [48]		●			2007
Wang et al. [71]		●	2		2009
Rind et al. [79]					2010
Faiola and Newlon [89]	16		●		2011
Gotz et al. [91]			2		2011
Gschwandtner et al. [92]	4	1			2011
Wongsuphasawat et al. [100]	●	1	2	10	2011
Zhang et al. [102]	6				2011
Alonso and McCormick [103]					2012
Sopan et al. [113]	3	●		●	2012
Wongsuphasawat and Gotz [115]	3	●		●	2012
Monroe et al. [122]	●				2013
Ramírez-Ramírez et al. [125]					2013
Borland et al. [128]					2014
Gotz and Stavropoulos [132]				12	2014
Kamaleswaran et al. [134]					2014
Malik et al. [136]		4		10	2014
Bernard et al. [144]	●		●		2015
Bernard et al. [143]	6				2015
Federico et al. [148]		5			2015
Klemm et al. [151]	3	3	2		2015
Glueck et al. [163]	2	2			2016
Jiang et al. [166]		●			2016
Kamaleswaran et al. [168]	4	4			2016
Loorak et al. [171]	5	5			2016
Ola and Sedig [177]					2016
Dabek et al. [189]		●			2017

Glueck et al. [193]		4	6		2017
Tong et al. [202]	4		1		2017
Tong et al. [201]	2		1		2017
Glueck et al. [209]	2	4			2018
Guo et al. [210]	1	1	3		2018
Tong et al. [218]			1		2018
Trivedi et al. [219]	9	9		9	2018
Alemzadeh et al. [222]					2019
Bernard et al. [224]	10	14		14	2019
Glicksberg et al. [228]					2019
Guo et al. [231]	3	3	2		2019
Kwon et al. [238]	2		1		2019
McNabb and Laramee [241]			3		2019
Sultanum et al. [248]	6	12		6	2019
Zhang et al. [251]	6	6			2019
Jin et al. [264]	2	7	2		2020
Kwon et al. [265]	9	9	1		2020
Wang et al. [298]	2	3	3		2021
<b>Total unique papers: 51</b>	30	26	17	8	

Table 2.15: **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.



## 2.6 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 organisations. As a response to the challenges stemming from healthcare data visualisation, we present a collection of open health data sets, and our methodology for searching for open healthcare data sets, 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 data sets. 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.

### 2.6.1 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 [175]. 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 increase manual labor. 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. Furthermore, an anonymisation 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 [247] 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 [175]. Over time, the secondary use of EHR data in supporting healthcare research is emerging and widely accepted worldwide, this in turn

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

Table 2.16: Keyword combinations used for discovering relevant healthcare data.

improves the quality control measures for collecting them [235].

*Data interoperability* is challenging, given there is no standard definition of an EHR, healthcare providers often develop their own format to support the clinical workflow [187]. The lack of standardised terminologies, 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 accessible EHR database [16, 167] and improving data validation and interoperability [67].

## 2.6.2 Healthcare Data Search Methodology

We focus on healthcare data sets that are openly accessible from a reputable data provider such as a nonprofit organisation, scientific 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 2.17. We check for conference associated events such as the annual IEEE Visualisation Contest dating back to 2004 [34], VAST challenges [207] and National NLP Clinical Challenges (n2c2) [246] for relevant data. We also use keyword combinations listed in Table 2.16 with data search engines [336, 335] and well-known government data portals [367, 321, 360, 362] to expand our results. We present 34 related healthcare data sets found in Table 2.17.

## 2.6.3 Healthcare Data Scope

The EHR data survey scope includes data sets 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 data set, or the most popular data sets if multiple data sets 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 fulfill some but not all criteria. We still include them as *context data sources* in our data source overview [Table 2.17](#).

We generally exclude data sets that require an access fee, with the exception of some candidates as context data sources. We generally exclude data sets that are accessible solely via project collaborations. We generally exclude data sets that are not archived in English. However, we do include some as context data sources (if they are high quality) in [Table 2.17](#) for interested readers, and describe them in [Section 2.6.5](#).

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

**Health IT Dashboard** [363] provides data sets on the adoption, utilisation, and performance of information technology in healthcare facilities sponsored by the US government, these data sets are excluded. **The VAST Challenge 2010 Mini Challenge 3** [84] provides a data set on genetic sequences for tracing the mutations of the Drafa virus. Each sequence of single molecules is coded as a single alphabet, therefore the data set does not contain any actual EHR information and is excluded. **The VAST Challenge 2011 Mini Challenge 1** [98] provides data containing posts collected from social media platforms for the identification of an epidemic outbreak, these data sets are excluded due to the lack of an EHR dimension.

### 2.6.4 Healthcare Data Sources Classification

We present a description of data sources in this section. [Table 2.17](#) 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 *specialised source* refers to data sets focusing on a single speciality or area of specialisation. The Human Mortality Database [361] provides multiple data sets specifically on all-cause mortality from over 50 countries or regions, therefore we classify it

Access	Specialised	Collection	Catalogue
Open Access	Human Mortality Database <sup>C</sup> [361] VAST Challenge 2010 Mini 2 <sup>[83]</sup> Project Tycho <sup>†</sup> [220] COVID-19 Dashboard <sup>C</sup> [278, 277, 257] The Scottish COVID-19 Response Consortium <sup>C</sup> [256]	UCI Machine Learning Repository [191] Public Health Wales [368] NHS Scotland Open Data [329] Data.gov.uk <sup>C</sup> [360] OpenDataNI [362] Global Health Data Exchange <sup>C</sup> [156] Big Cities Health Coalition [327] NHS England [325] Public Health England [332] City Health Dashboard [230]	FAIRsharing <sup>C</sup> [333] Data.gov <sup>C</sup> [321] HealthData.gov <sup>C</sup> [338] European Data Portal <sup>C</sup> [367] Maelstrom Catalogue [358] re3data <sup>C</sup> [356] COVID-19 Open Research Data set Challenge <sup>C</sup> [263]
Verification Protocol	National NLP Clinical Challenges [246] MIMIC-III [167]	Study of Health in Pomerania <sup>††</sup> [20] PhysioNet <sup>†</sup> [16] SAIL Databank <sup>†</sup> [67] Health Data Research Innovation Gateway <sup>†</sup> <sup>C</sup> [249]	
Fee and Verification Protocol		Rotterdam Study <sup>†</sup> [326] GIANTT <sup>††</sup> [334] TRAILS <sup>††</sup> [366] LifeLines Biobank <sup>C</sup> [322] UK Biobank [323] SEER Program <sup>†</sup> [344]	

Table 2.17: **Data source table:** Data sources ordered by the year of establishment. See the detailed description of focus data sources in Section 2.6.5. Green highlights context data sources. <sup>C</sup>Contains COVID-19 data. <sup>†</sup>Registration required for open access. <sup>‡</sup>Partially open access. <sup>††</sup>Free access for project collaborators, paid access for noncollaborators. <sup>¶</sup>Free access for project collaborators, no access for noncollaborators. <sup>††</sup>Data is not archived in English.

as a specialised focus data source.

A *collection source* provides access to multiple data sets from different specialities, such as the UCI Machine Learning Repository [191], 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) [356] is a catalogue source that hosts over 2,000 scientific data sets, each comes with a comprehensive description and a link pointing to its homepage.

## 2.6.5 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 [127]:

- 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

### Specialised Healthcare Data Sources

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

**Human Mortality Database** began as a collaborative project in 2000 [361],

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** [83], 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 [220], 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 data sets 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 [278, 277, 257], as a real-time visualisation 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, and 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 [254].

## Collection Healthcare Data Sources

This section describes focus data sources that provide access to multiple data sets from different specialities.

**UCI Machine Learning Repository** was created by David Aha and fellow graduate students at the 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 data sets, 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)[368], **NHS Scotland Open Data** (2009)[329], **The Government Digital Service** (2011)[360], **OpenDataNI** (2012)[362], **Public Health England** (2017)[332] and **NHS England** (2017)[325]. Example data sets hosted on these portals include mortality rates from cancer, liver, cardiovascular diseases, and more.

**Big Cities Health Coalition** [327] is a forum founded in 2014, that 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** [156] 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.

## 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** [333] started in 2007 as a community-driven registry providing descriptions of standards, databases, and data policies. Data sets can be published on FAIRsharing to increase 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** [321] and **HealthData.gov** [338] started

offering links to data sets in 2011, to ensure compliance with relevant Open Data Policy and to promote research and innovation. Public entities ranging from federal agencies to local government departments collected over 200,000 data sets, including popular healthcare data on cancer, diabetes, and hypertension.

The **European Data Portal** [367] 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 data sets including HIV-related, norovirus, and cancer are available.

**Maelstrom Catalogue** [358] 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** [356] 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 Data set Challenge** [263] is a challenge launched in 2020 by the Allen Institute for Artificial Intelligence on Kaggle, an on-line 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 anonymised patient data, as a part of the submission requirements.

## Context Healthcare Data Sources

A context healthcare data source refers to a data source that does not fulfill all criteria listed in [Section 2.6.3](#), but we include and describe some high quality sources here for interested readers.

**UK Biobank** [323] 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** [60, 322] 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)** [366] 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 conditions, 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** [326] 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 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** [67] was established in the UK in 2006. It allows external researchers to access billions of EHRs on data sets such as outpatient, critical care, and primary GP care in the UK. Access to additional restricted data sets 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)** [20, 99] 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 Analyze Type 2 Diabetes Treatment (GIANTT)** [334] 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 practises. 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.

## 2.7 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 2.18 summarises 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 utilisation. 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.

Literature	Challenges									Year	
	Scalability	Evaluation	Screen space	Interoperability	Usability	Interaction	Dimensionality	Uncertainty	Clustering		Access
Plaisant et al.[11]											1998
Horn et al.[19]											2001
Bade et al.[30]											2004
Goren-Bar et al.[33]											2004
Hinum et al.[39]											2005
Fails et al.[45]											2006
Bui et al.[48]											2007
Wang et al.[71]											2009
Rind et al.[79]											2010
Faiola and Newlon[89]											2011
Gotz et al.[91]											2011
Gschwandtner et al.[92]											2011
Wongsuphasawat et al.[100]											2011
Zhang et al.[102]											2011
Alonso and McCormick[103]											2012
Sopan et al.[113]											2012
Wongsuphasawat and Gotz[115]											2012
Monroe et al.[122]											2013
Ramírez-Ramírez et al.[125]											2013
Borland et al.[128]											2014
Gotz and Stavropoulos[132]											2014
Kamaleswaran et al.[134]											2014
Malik et al.[136]											2014
Bernard et al.[144]											2015
Bernard et al.[143]											2015
Federico et al.[148]											2015
Klemm et al.[151]											2015
Glueck et al.[163]											2016
Jiang et al.[166]											2016
Kamaleswaran et al.[168]											2016

table continued on next page ...

... continued

Literature	Challenges										Year
	Scalability	Evaluation	Screen space	Interoperability	Usability	Interaction	Dimensionality	Uncertainty	Clustering	Access	
Loorak et al.[171]											2016
Ola and Sedig[177]											2016
Dabek et al.[189]											2017
Glueck et al.[193]											2017
Tong et al.[202]											2017
Tong et al.[201]											2017
Glueck et al.[209]											2018
Guo et al.[210]											2018
Tong et al.[218]											2018
Trivedi et al.[219]											2018
Alemzadeh et al.[222]											2019
Bernard et al.[224]											2019
Glicksberg et al.[228]											2019
Guo et al.[231]											2019
Kwon et al.[238]											2019
McNabb and Laramee[241]											2019
Sultanum et al.[248]											2019
Zhang et al.[251]											2019
Jin et al.[264]											2020
Kwon et al.[265]											2020
Wang et al.[298]											2021
<b>Total unique papers: 51</b>	24	15	12	10	13	9	7	7	4	4	

Table 2.18: **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. **Green** highlights context papers. We use 1-2 words to represent these challenges in the table header, and describe them in detail in [Section 2.7](#).

**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 [46]. 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 [113]. From the table, we can see this has been a persistent theme.

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

In-depth **evaluation (14 papers, 29%)** and validation including quantitative studies, qualitative studies, and validation are reported 14 times as the second most popular future research direction. An in-depth evaluation and validation help to reveal the weakness and potential improvements for the system. We examine and describe the evaluation techniques adopted by the literature in [Section 2.5](#). 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 is a prominent theme in recent years.

Limited **screen space (12 papers, 24%)** constrains the content visualised and reduces the effectiveness of an EHR system [228]. As the probability of using multiple views increases in EHR Vis systems, we categorise this challenge as a domain-specific one. Features with less significance are often hidden to make space for others [224]. This may result in over-simplification and missing potential insights [122]. 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. [Section 2.7](#) 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 [175] and is reported 10 times as a future challenge. This increases the difficulty for researchers to incorporate data from heterogeneous sources in varying formats [177]. Although [Table 2.3](#) indicates that some papers focus on the same UMLS terms, these EHR Vis systems are built specifically for their given data sets 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 the same topics and adopting a common terminology standard such as the UMLS. [Section 2.7](#)

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 [71, 169]. 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 contain missing or incorrect values, this requires further investigation during data collection and preprocessing [178].

**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 [175]. We attempt to address this challenge here in Section 2.6. 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 visualisation and medical experts: in Section 2.5, we find 59% of the papers choose to collaborate with medical experts, who also provide EHR data for visualisation researchers.

Literature	Visualisation Techniques																					Year
	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	
Plaisant et al.[11]																						1998
Horn et al.[19]																						2001
Bade et al.[30]																						2004
Goren-Bar et al.[33]																						2004
Hinum et al.[39]																						2005
Fails et al.[45]																						2006
Bui et al.[48]																						2007
Wang et al.[71]																						2009
Rind et al.[79]																						2010
Faiola and Newlon[89]																						2011
Gotz et al.[91]																						2011
Gschwandtner et al.[92]																						2011
Wongsuphasawat et al.[100]																						2011

table continued on next page ...

...continued

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
Zhang et al.[102]																		●				2011
Alonso and McCormick[103]							●		●							●						2012
Sopan et al.[113]					●		●									●						2012
Wongsuphasawat and Gotz[115]														●								2012
Monroe et al.[122]							●															2013
Ramírez-Ramírez et al.[125]									●							●						2013
Borland et al.[128]				●								●										2014
Gotz and Stavropoulos[132]			●				●	●	●				●	●								2014
Kamaleswaran et al.[134]							●															2014
Malik et al.[136]							●									●		●				2014
Bernard et al.[144]																●						2015
Bernard et al.[143]	●	●						●								●						2015
Federico et al.[148]							●									●				●		2015
Klemm et al.[151]							●															2015
Glueck et al.[163]							●									●		●				2016
Jiang et al.[166]	●					●										●						2016
Kamaleswaran et al.[168]		●				●					●						●					2016
Loorak et al.[171]								●			●	●				●						2016
Ola and Sedig[177]			●		●	●	●	●		●			●	●		●					●	2016
Dabek et al.[189]							●				●					●		●	●	●		2017
Glueck et al.[193]							●	●		●								●				2017
Tong et al.[202]				●												●					●	2017
Tong et al.[201]				●												●					●	2017
Glueck et al.[209]							●								●	●		●	●			2018
Guo et al.[210]							●									●			●		●	2018
Tong et al.[218]				●																		2018
Trivedi et al.[219]											●									●		2018
Alemzadeh et al.[222]					●						●	●				●						2019
Bernard et al.[224]		●					●	●								●						2019
Glicksberg et al.[228]																●						2019
Guo et al.[231]							●									●					●	2019
Kwon et al.[238]	●						●	●							●							2019
McNabb and Laramée[241]	●						●									●						2019
Sultanum et al.[248]							●				●							●	●			2019
Zhang et al.[251]		●					●									●						2019
Jin et al.[264]							●						●			●						2020
Kwon et al.[265]					●		●					●	●			●	●					2020
Wang et al.[298]							●				●				●							2021
<b>Total unique paper: 51</b>	4	4	3	3	4	3	25	6	8	3	8	4	4	3	3	30	2	6	11	4	6	

Table 2.19: Visualisation techniques applied in the literature. We follow the classification of visualisation techniques by Keim[26], and categorise bar chart, line chart and pie chart as standard 2d display. **Green** highlights context papers. **●** indicates the technique is applied in the literature. **●** indicates a customised variant of the technique is applied in the literature.

**More advanced visual designs:** Table 2.19 shows an overview of visualisation techniques applied in all papers included in this STAR. We observe that standard 2D

displays and glyphs are the most popular techniques among 21 techniques found across all EHR Vis systems. This implies that using advanced visual techniques to mitigate scalability challenges brought by EHR data dimensionality, remains understudied.

## 2.8 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 Visualisation and Visual Analytics. We investigate some of the most commonly used terminologies in the field and categorise 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 utilises 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 data sets, which 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.

By providing a comprehensive survey of the state-of-the-art in EHR Vis, including the classification of literature and identification of multidisciplinary themes, the chapter establishes key challenges and opportunities within the field. These insights guide the development of novel visualisation techniques and evaluation methods in the following chapters. Specifically, the categorisation of research themes and the analysis of EHR data types, such as text data explored in [Chapter 3](#), geospatial in [Chapter 4](#), and long time series visualisations in [Chapter 5](#).

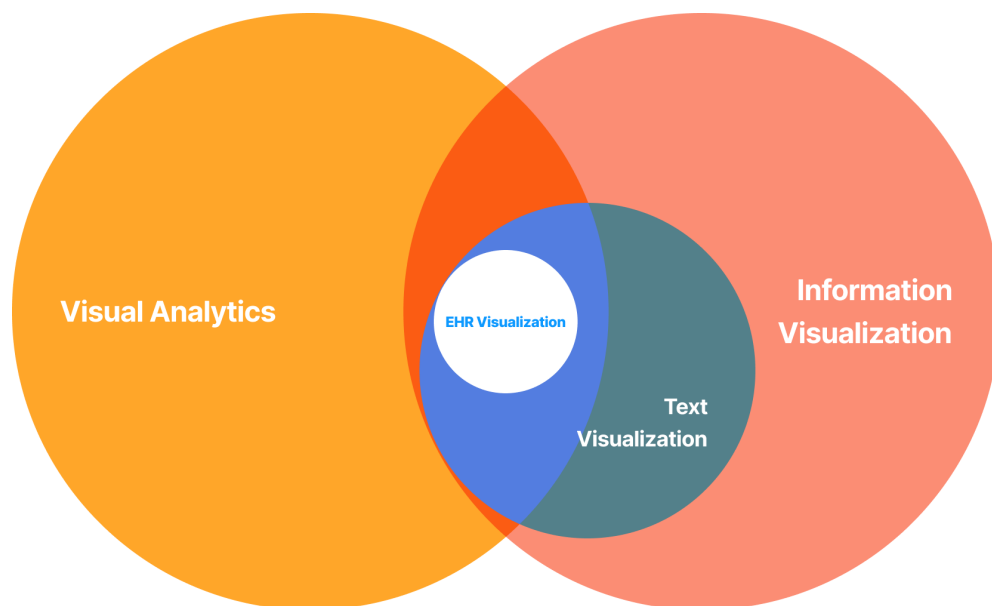




## Chapter 3

# LetterVis: a Letter-space View of Clinic Letters

Wang, Q., Laramée, R. S., Lacey, A., & Pickrell, W. O. (2021). LetterVis: A letter-space view of clinic letters. *The Visual Computer*, 37(9–11), 2643–2656. <https://doi.org/10.1007/s00371-021-02171-w> [298]



*“Visualisation gives you answers to questions you didn’t know you had.”*

– Ben Shneiderman, the Father of Treemap (1947 - present)

This chapter is based on the publication in *The Visual Computer* [298]. The inspiration for this work stems from the need to understand the content of clinic letters, which are the primary means of communication between healthcare professionals across different departments. The void of visualisation tools for clinic letters, identified in our survey presented in [Chapter 2](#), motivated us to develop LetterVis, a visualisation tool that offers a letter-space view of clinic letters.

## Contents

---

<b>3.1</b>	<b>Introduction</b>	<b>73</b>
3.1.1	Motivation	73
3.1.2	Contribution	74
<b>3.2</b>	<b>Related Work</b>	<b>75</b>
<b>3.3</b>	<b>Data Description</b>	<b>77</b>
<b>3.4</b>	<b>Application Design Methodology</b>	<b>78</b>
3.4.1	Informing the Initial Design	78
3.4.2	Informing Further Software Iterations	79
3.4.3	User Requirements and Design Goals	79
3.4.4	Tasks	80
<b>3.5</b>	<b>LetterVis for Visualisation of EHR Letters</b>	<b>81</b>
3.5.1	Initial Prototype	82
3.5.2	Three Levels of Letter Abstraction	84
3.5.3	Advanced Visual Filtering and Selection	86
<b>3.6</b>	<b>Evaluation</b>	<b>90</b>
3.6.1	Case Studies	90
3.6.2	Domain Expert Feedback	94
3.6.3	Domain Expert Review	96
<b>3.7</b>	<b>Limitations and Future Work</b>	<b>98</b>
<b>3.8</b>	<b>Conclusions</b>	<b>98</b>
<b>3.9</b>	<b>Appendix</b>	<b>100</b>
3.9.1	List of Expert Sessions	100
3.9.2	List of Interview Questions	100

---

## 3.1 Introduction

Having a comprehensive understanding of the EHR Vis field, we begin to develop our first EHR Vis system.

Routinely collected Electronic Health Records (EHR) such as clinic letters contain important information such as demographics, past and present prescriptions and previous check-ups, all of which are valuable in answering clinical research questions. These letters are often stored in a free-text format by clinicians with different writing styles, thus making the extraction of critical information for further analysis a time-consuming and error-prone process. Even with the assistance of machine learning and modern statistical methods, effectiveness remains limited, let alone the challenges associated with transparency, replicability and ethics [282].

Visual analytics (VA) and visualisation often involve real-time human observation and intervention. VA has great potential to support clinical decision-making and inform further research under close scrutiny to ensure both quality and transparency. Interactive visual designs can efficiently reduce visual clutter to cope with the exploding data volumes, supporting quantitative as well as qualitative analysis in an interpretable and explainable manner [280]. We propose LetterVis, an interactive letter-space visualisation tool specifically designed to enable efficient exploration and analysis of clinic letters. By letter-space, we mean the standard coordinate system used by clinicians to write clinic letters, e.g. A4 space. See [Section 3.5](#) for more details. Our collaboration with domain experts informs a novel design that utilises the information-rich clinic letters to assist in hypothesis generation and verification, via the human visual perceptual system. We focus on epilepsy in our case studies, but our tool can be easily generalised to support all content in the form of letters.

### 3.1.1 Motivation

The motivation behind our study is to help EHR researchers explore and analyse AED (antiepileptic drug) co-prescriptions in unstructured EHR letter data to identify pattern patterns and outliers, and also to provide an overview, filtering and selection, and analysis options. From our interviews with healthcare data analyst experts, they report

that their recent adoption of advanced visual designs and VA for analysis of AED co-prescriptions yields promising outcomes. However, they point out that usability is an obstacle to furthering the analysis of AED co-prescriptions, as they often encounter a steep learning curve. This in turn may result in more human errors.

### 3.1.2 Contribution

The development of LetterVis represents a significant contribution to the field of EHR Vis, specifically addressing the challenges associated with unstructured clinical text. Unlike traditional approaches that rely on static displays or simple keyword searches, LetterVis provides a dynamic, multilevel abstraction framework that allows users to explore clinical letters interactively. Existing methods, as detailed in [Chapter 2](#), often lack the ability to balance granularity with an overarching view of the data set, making it difficult to extract actionable insights from fragmented and text-heavy data.

Our contributions include:

- A novel letter-space visualisation tool that leverages natural language processing (NLP) techniques to support the exploration of unstructured clinical text in a structured manner,
- Novel and customised visual designs to identify and verify patterns and outliers in a cohort of patients,
- Dynamic analysis and comparison of antiepileptic drug (AED) co-prescriptions through multiple coordinated visual layouts,
- Three replicable case studies to demonstrate LetterVis' ability to support hypothesis verification.

Additionally, the iterative design process, conducted in collaboration with EHR domain experts, ensures that the tool meets the practical needs of its users. LetterVis not only enhances the accessibility and usability of textual clinical data but also establishes a foundation for integrating text-based insights into broader healthcare visualisation systems, as explored in subsequent chapters of this thesis.

## 3.2 Related Work

The specific related work to ours falls into the subcategories: EHR Vis and Natural Language Processing, and Advanced Query Interfaces. Our work advances state-of-the-art in EHR Vis and Advanced Query Interfaces (not NLP). We simply use some NLP preprocessing to enhance visualisation. Based on the comprehensive survey described in [Chapter 2](#), we quickly narrow down the scope of related work to the following publications.

Bernard et al. [\[145\]](#) build a visual-interactive system that enables physicians to train models for prostate cancer identification. Glueck et al. present a trilogy of visual analysis tools for phenotype comparison: PhenoBlocks [\[163\]](#) with a novel differential hierarchy comparison algorithm accompanied by a customised sunburst radial hierarchy layout, PhenoStacks [\[193\]](#) with a novel topology simplification algorithm to eliminate duplicates, and incorporates natural language queries for searching, and PhenoLines [\[209\]](#) adds the support for the visualisation of the temporal evolution of phenotypes.

Machine Learning (ML) literature shows an emerging trend in the field of EHR Vis. Kwon et al. [\[238\]](#) apply recurrent neural networks (RNN) to produce predictions based on the temporal dimension in EHRs. Jin et al. [\[264\]](#) leverage machine learning techniques to assist the preprocessing and visualisation of EHR data. RegressionExplorer [\[226\]](#) provides an interactive visual interface for clinical biostatisticians to verify and improve their models through visualisations. The black box nature of these approaches impedes the justification of clinical decisions suggested by the models [\[282\]](#). Our approach on the other hand, provides a transparent process to lower the explanatory burden. Also, the focus of our work is not ML.

Natural Language Processing (NLP) is defined by Liddy as “*a theoretically motivated range of computational techniques to analyse and represent 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*” [\[29\]](#). NLP plays a significant role in the visualisation and visual analytics of EHR data archived as free text. EHR-NLP approaches usually include the use of existing tools, customised classification algorithms, and curation-based extraction [\[236\]](#).

Through our domain expert partners, we learn that GATE, an open source text

analysis tool that supports text mining of biomedical documents via natural language processing (NLP), developed by Cunningham et al. [120], is one of the most popular choices when it comes to EHR data preprocessing. GATE is capable of extracting structured information from unstructured free text, e.g. clinician’s notes and discharge letters. However, it lacks interactive visualisation features to assist advanced analysis.

Zhang et al. incorporate NLP algorithms in their AnamneVis [102] to extract structured medical information from doctor-patient dialogs and medical reports to assist in the visualisation of patient medical history. Trivedi et al. [219] introduce NLPReViz which interactively trains NLP models to classify information extracted from clinical records. Sultanum et al. present Doccurate [248] which provides an accurate and sufficient overview for individual patients based on user-supplied extraction rules. The focus of our work is not NLP itself, but rather extracting a structured visual representation of the data hidden inside letters with the help of NLP.

Stubbs et al. introduce Sim•TwentyFive [114], an intuitive visual querying interface for decision support in psychiatric intensive care units. The system is specifically designed to work with small screen devices.

Event Sequence Simplification (ESS) is one of the popular techniques used to support the visualisation of high-dimensional temporal event sequence data [8]. Previous work focuses on the aggregation of events in EHR data while preserving potential insight [123, 137] and utilises various visual designs to provide a clear overview [100] or a simplified comparison of multiple patients [45, 116]. In DecisionFlow [132], users construct multiconstraint queries to prepare EHR data for further analysis.

Different healthcare facilities adopt different medical terminology standards, such as ICD-10 [340], Read Codes [357], SNOMED-CT [339], and the UMLS we have investigated in star-subsubsec:Adopting a Medical Terminology Standard. Glicksberg et al. describe PatientExploreR [228], an interactive querying interface that enables physicians to quickly search and filter EHRs collected from multiple sources written in varying medical terminology standards.

The key difference between our work and previous work is the introduction, development, and evaluation of letter-space, and its application to clinic letters. We work closely with health data analysts to curate a specific list of extraction rules for pro-

cessing epilepsy-related EHRs. To facilitate the analysis of the result, we incorporate interactive visual designs along with an advanced query interface that is compatible with Apache Lucene [149], a text search engine library that is known for its flexible and efficient search algorithms. The library came to our attention after one of our expert health data analyst partners' recommendations.

### 3.3 Data Description

Our data includes 200 clinic letters written by neurology-specialist clinicians and provided by our healthcare data analyst expert partners in Word format. Each letter represents a single patient visit to a neurologist. A typical letter contains identifiable information including patient name, age, gender, address, NHS number, AED prescriptions (past and present), symptoms, diagnosis, and other health-related information. Due to the sensitive nature of EHR data, the letters are manually anonymised by clinicians with patient identifiable information as well as other potentially identifiable information manually replaced with similar but fictional text.

We first extracted all text from Word files, the length of letters varied from 28-133 lines with an average of 98 characters per line. We then preprocessed the letters using NLP with a list of curated extraction rules provided by domain experts to extract 12 text data categories. Our extraction rules started with numerical data as a proof of concept. This was then expanded to include antiepileptic prescription information based on our health data analysts' feedback. We also extracted metadata for these categories, such as the position and length of the matching text. The extracted data was then used to generate visualisations matching the data samples' position in the original letters.

In later iterations of the software development, we received a list of 26 generic antiepileptic drug (AED) names together with 24 equivalent trade names from our domain expert partners for the purpose of exploring patterns in drug co-prescriptions and effective combinations. Each AED colour was assigned from the colormap shown in Figure 3.3A. We develop the matrix view (Section 3.5.3) and drug chain view (Section 3.5.3) based on this list and the user requirements in Section 3.4, with an additional

Category	Description	Items Extracted
Phone	Phone numbers	38
Postcode	Postcode	352
Date	Date in various formats	728
Time	Time in various formats	723
Hospital & NHS Number	A unique 10-digit number	79
Date of Birth	Patient’s date of birth	202
Age	Patient’s age	141
Measurement	Patient’s measurements, eg. weight and height	20
Drug	Name of medicines	884
Dosage	Dosage of medicines	765
Frequency	Frequency of medicines	315
Other	Other potentially interesting numerical values	1,442

Table 3.1: 12 text data categories extracted from our collection of letters, with a brief description and the total number of items extracted for each category.

pre-defined colour legend for each AED shown in [Figure 3.3C](#).

## 3.4 Application Design Methodology

We describe our collaboration with three health data analyst experts in this section.

We collaborate closely with a consultant neurologist (E1) from the UK National Health Service (NHS), and a lecturer in Natural Language Processing who is also a senior health informatics research analyst in epilepsy-related research (E2) from a UK University. We also interview a health data scientist (E3) from a UK University Medical School during the initial design stage. Data for this application is provided by E1 and E2, described in [Section 3.3](#). See [Table 3.4](#) for a list of interview and feedback sessions. Throughout the entire development process, we also exchanged over 30 emails with E1 and E2 to discuss the design and implementation of LetterVis.

### 3.4.1 Informing the Initial Design

Our initial design was informed by interviewing three health data analyst experts in EHR analysis. We follow the guidance of Hogan et al. [\[165\]](#) and constructed a set of interview questions involving 14 structured, semi-structured and open-ended questions. See [Section 3.9.2](#). These questions were carefully crafted to ensure comprehensive exploration of the topic, with many being open-ended to encourage participants to share detailed insights and elaborate on their experiences. The open-ended nature of these questions was instrumental in gathering nuanced information, fostering a deeper



understanding of the domain. All interviews were audio recorded for post-analysis with consent, ensuring that no valuable insights were missed.

We then analysed the domain requirements to guide the development.

### 3.4.2 Informing Further Software Iterations

Over the course of development spanning over 12 months, we consulted E1 and E2 for feedback. E3 did not participate in the development as her specialisation in injury prevention is not related to the letters on epilepsy that we worked on. We presented intermediate visualisation prototypes to E1 and E2 in four separate feedback sessions. Each session lasted around 65 minutes and was recorded for post-analysis. We describe this feedback in detail in [Section 3.6.2](#).

Our work can be easily extended and generalised to support other areas where letters are used systematically for communication, either currently or historically, for example, in the legal profession.

### 3.4.3 User Requirements and Design Goals

LetterVis was developed in collaboration with three health data analyst experts in clinic letter analysis. The following requirements are gathered from interviews and feedback sessions:

- R1** An interactive tool that facilitates the exploration of EHR free text data,
- R2** Software that supports the identification of patterns and outliers with respect to AED co-prescriptions in clinic letters,
- R3** A design that supports the identification and exploration of AED co-prescriptions,
- R4** An interface that supports analysing several clinic letters simultaneously,
- R5** Support for cross-referencing and linking visual representations with original letters,
- R6** A tool that is compatible with experts' existing analytical workflow by supporting EHRs in JSON format.

Throughout the development process, we identified additional requirements from feedback:

- R7** A query interface that is compatible with Apache Lucene syntax,

**R8** An interface that conveys AED prescription evolution.

### 3.4.4 Tasks

Brehmer and Munzner’s multilevel typology of visualisation tasks [118] provides guidance on classifying and describing our visualisation tasks.

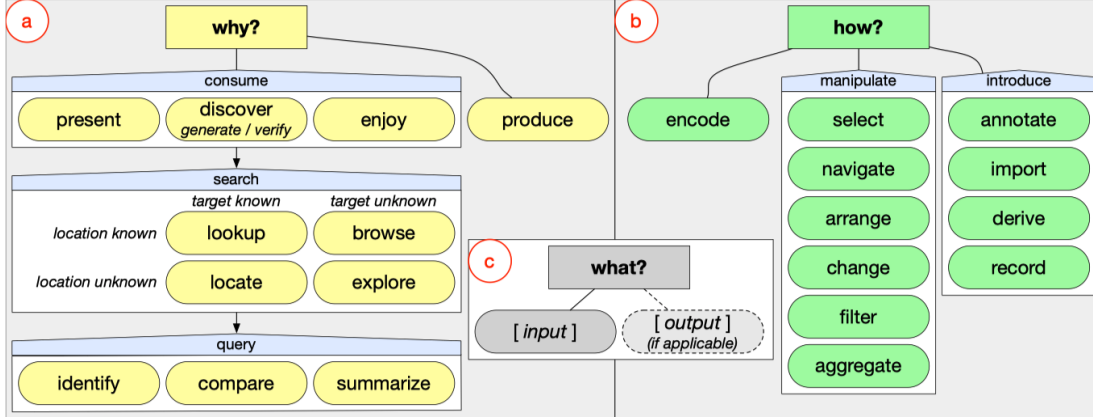


Figure 3.1: The multilevel typology of abstract visualisation tasks by Brehmer and Munzner. Figure reproduced from Brehmer and Munzner [118].

Based on the topology, we derive six main tasks to meet the requirements above [118]:

- T1** *Present* an overview of important text data with abstraction of user-chosen data-of-interest, that enables the user to *explore*, *identify* and *compare* patterns and outliers (**R1, R2, R4**). [present → explore → identify/compare]
- T2** A coordinated visual interface that *presents* multiple levels of abstracted views to support the *exploration* of letters and *identification* of patterns and outliers (**R1, R2**). [present → explore → identify]
- T3** A combination of customised visual designs for visualising AED co-prescriptions and prescription progression, the interface enables the user to *lookup* and *compare* different letters and *select*, *arrange*, *change* and *filter* based on AED co-prescriptions (**R3, R4, R5, R8**). [lookup → compare → select/arrange/change/-filter]
- T4** Develop a visual query interface that is compatible with Apache Lucene syntax to support the existing analytical workflow which enables the user to *identify* outliers (**R6, R7**). [identify]

**T5** Provide a range of interactive user options and their combinations, including *filtering* and *selection*, to support tasks **T1**, **T2** and **T3**. [select/filter]

**T6** Provide a history of queries that supports the retrospective analysis through undo and redo functions. [record]

Our design follows the Visual Information-Seeking Mantra, “*overview first, zoom and filter, then details on demand*” [9], as a starting point. Shneiderman further proposes three essential tasks: *Relate* to view relationships between items, *History* to keep a list of actions performed and provide undo and redo functions, *Extract* to enable extraction of sub-collections. We believe the tasks established above are generic in most free text analysis projects, therefore our work can be extended to support other domains.

### 3.5 LetterVis for Visualisation of EHR Letters

Valuable patient information is recorded and exchanged in the form of clinic letters to deliver quality patient care. Although letter text is usually described as unstructured, our basic hypothesis and motivation is based on the implicit knowledge and hence structure hidden in letters. For example, postcodes do not appear at random positions in the letters. Their position is consistently in the top-left in letter-space. We believe that the position of numerical data in letter-space can provide important clues about the context hidden inside the unstructured text, likewise for drugs and prescriptions (For example, see [Figure 3.5](#)). Another reason we focus on letter-space is because this is the space that clinicians and EHR analysts operate in and are used to. Any of the unfamiliar visual designs that we develop can be linked back to the familiar letter-space to facilitate interpretation by the analysts and/or clinicians that write them in the first place. This is crucial for any interdisciplinary project. Also, position of data on a page is important because it can reveal outliers (**R2**, **T1**). The order in which drugs appear is not random. The choice and order of prescriptions reflect important medical and pharmaceutical expertise held by clinicians (For example, see [Figure 3.4](#)). Our visual approaches extract this information and leverage it to facilitate the exploration and analysis of clinic letters.

We construct a letter-space by deriving the width (x-axis) from the clinic letters.

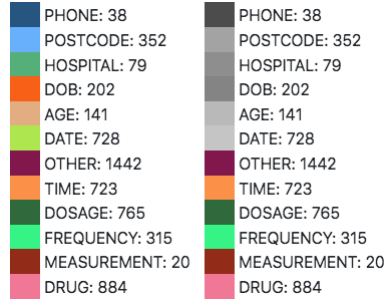


Figure 3.2: An illustration of the colour legend for 12 numerical categories. Clicking on categories will render the corresponding centroids and individual samples in the global, thumbnail and focus views as context (in greyscale). Each data category is followed by the number of matches found in the uploaded letters.

The width is the average length of text, approximately 98 characters, derived from the full collection of letter bodies. We then use the standard letter aspect ratio to calculate the height (y-axis) for depicting a letter-space in all three letter abstraction views described in [Section 3.5.2](#).

Our letter-space approach facilitates the real-time exploration of clinic letters, streamlining and enhancing the decision-making process. Designed to align with clinicians’ existing analytical workflows, this approach is intuitive and minimises cognitive load. In this section, we provide a detailed description of the five customised visual interfaces we propose. Additionally, we introduce features that support visual queries and sorting options, further improving usability and efficiency.

### 3.5.1 Initial Prototype

An initial prototype was developed after the first round of interviews with our domain expert partners. The prototype included some basic features to visualise numerical data categories extracted from the clinic letters and a colour legend to represent the data categories. These features were later refined and expanded based on feedback from our domain expert partners.

#### Visualisation of Numerical Values

In the first few iterations of LetterVis, we focused on visualising numerical data as proof of concept.

Specifically, we focused on both the type of numerical data and its position in



Figure 3.3: This figure shows an overview of LetterVis. We also provide a detailed description of visual designs and elements in their respective sections. (A) shows the user options for searching, rendering, and sorting (Section 3.5.3). (B) illustrates the matrix view based on AED co-occurrences in the data set, user-chosen cells are highlighted with a sequence number, which corresponds to their position in the history of queries (Section 3.5.3). (C) The drug chain layout returned by the queries, user-chosen AEDs are shown in colour (Section 3.5.3). (D) depicts an overview of all super-imposed letters and their search result centroids in the data set (Section 3.5.2). (E) shows individual thumbnails of letters with local search result centroids joined by edges (Section 3.5.2). (F) contains a detailed view of the letter in focus, lines without data-of-interest are collapsed by default (Section 3.5.2).

letter-space. We extracted and classified all numerical values (Figure 3.3A left) using customised NLP extraction rules based on regular expression, and visualised the distribution of values to depict clusters in letter-space. As we anticipated, extraction rules constantly evolve during our collaborative development lifecycle to meet new requirements proposed by our domain experts.

We decided to use regular expressions for their flexibility to quickly prototype and refine our software. The numerical categories were used to support analytical tasks in all case studies in Section 3.6.1. We chose this approach as a starting point to demonstrate

out idea to the experts in order to get feedback and inform future software features. One limitation of this categorisation of data quickly identified by our domain expert partners was the ability to query data by keyword, which is essential to their analytical workflow.

### Color Legend Interaction

We use Colorgorical [194] to create a discriminable and aesthetically guided colormap to represent the 12 data categories extracted. The legend is shown in Figure 3.2 and Figure 3.3A left. Any legend component can be clicked or dragged to the search bar to initiate a category search. Clicking on a legend item toggles the rendering (in-focus) of the corresponding category and individual samples in all three abstraction views, as shown in Figure 3.5. Figure 3.3A bottom shows a colour legend for AEDs. On-mouse-over displays the AED name. This part of the interface can be hidden if the user would like to reduce the complexity of the interface.

### 3.5.2 Three Levels of Letter Abstraction

We provide three different views to represent the abstraction of letters from a top-down perspective. The first level enables exploration at the cohort level for global analysis (Figure 3.3D). The second level represents the abstraction of each individual letter by juxtaposition for closer observation (Figure 3.3E). The third level links detailed visual elements with the original letter (Figure 3.3F). This approach enables the user to explore the letters at different levels to identify patterns and outliers in a cohort of patients. All views are linked and coordinated, supported by interactive user options.

### Visual Elements in the Global and Thumbnail Views

In the global and thumbnail views, we introduce three visual elements:

- *Centroid*: represents the arithmetic mean position of each text data category (numerical, AED and search term-based) in letter-space
- *Individual sample*: represents an individual text data sample in letter-space
- *Edge*: connects a centroid to its individual samples in the same data category

## Global View

The global view ([Figure 3.3D](#)) is the first and highest level of abstraction that shows all search term samples in letter-space and their corresponding category centroids extracted from the data set in one superimposed letter (**T1**). This superposition approach enables the comparison of search terms in all of the letters simultaneously, which is otherwise difficult or even impossible through juxtaposition or explicit alignment. Using juxtaposition or explicit alignment to obtain an overview or make comparisons is ineffective in this case. Clicking on a centroid triggers the rendering of edges to the corresponding individual search term samples ([Figure 3.5](#)). For example, [Figure 3.5](#) top left shows the centroids of each data category listed in the colour legend in the global view of letter-space (left). The greyscale points are rendered as context. They are the positions of the original search data samples in letter space. Selecting a single search dimension, e.g., Drugs, causes edges to be rendered from the search data dimension centroid to individual samples ([Figure 3.5](#) bottom left). We render edges because they convey the area covered by a category of values in letter space. Also, search term samples located further from the centroid often have a higher chance of depicting outliers. A convex hull could have been used, however, it does not show the variation and density of the original search term samples. Clicking on an individual sample in the global view shows the corresponding letter in the focus view (**T2**). On-mouse-over details are provided for every visual element.

## Thumbnail View

As the second level of abstraction (**T1**, **T2**), each thumbnail juxtaposed in the thumbnail view ([Figure 3.3E](#)) represents an individual letter. Similarly to the global view, a user clicks on a centroid to show each connection to individual samples of the query. Clicking on the title shows the corresponding letter in the focus view.

On-mouse-over information shows the original data for every visual element.

## Focus View

The focus view is the third level of abstraction (**T2**) that shows a summarised version of letters ([Figure 3.3F](#)). Individual samples are highlighted. Lines with no text data of

interest are collapsed by default and can be expanded interactively via clicking on any arrow glyph. At any stage, clicking on the ‘View Document’ button brings the original letter, with all individual samples highlighted, into focus. An example is shown in [Figure 3.6](#).

### 3.5.3 Advanced Visual Filtering and Selection

LetterVis facilitates data exploration via a visual querying interface combined with rendering and sorting options. We also color-mapped visual elements and provide further interaction via colour legends.

#### Matrix View

We include a co-occurrence matrix specifically for visualising AED co-prescriptions, as a special requirement requested by our domain expert partners to support the exploration of common and unusual AED co-prescriptions (**T3**). Co-occurring AEDs appear as colour-coded cells where a row and a column intersect in the matrix view ([Figure 3.3B](#)). We extract AEDs in real-time. By exploring common and unusual AED co-prescriptions visually, they can potentially reduce the number of trials needed for finding optimal AED co-prescriptions for patients. The matrix view ([Figure 3.3B](#)) is automatically rendered when letters are loaded. Popular co-prescriptions are trivially observed when the matrix view is sorted by co-occurrence frequency. Co-prescriptions with a higher frequency are also rendered in more distinct colours than others. Hovering over a cell initiates an arrow from the y-axis to the x-axis connecting the corresponding pair of AEDs and also highlighting both AEDs (**T5**). See [Figure 3.3B](#).

#### Drug Chain View

The matrix view only indicates the co-occurrence of pairwise drugs. When a query involves multiple AEDs, the drug chain view ([Figure 3.3C](#)) is rendered. In the drug chain view, blocks representing multiple AEDs are linked in order of appearance in the corresponding letter. This view aims to provide a visual representation of prescription progression and may unveil unique insight into epilepsy progression as well(**T3**). User-chosen AEDs are shown in colour (focus), while the remaining AEDs are rendered as



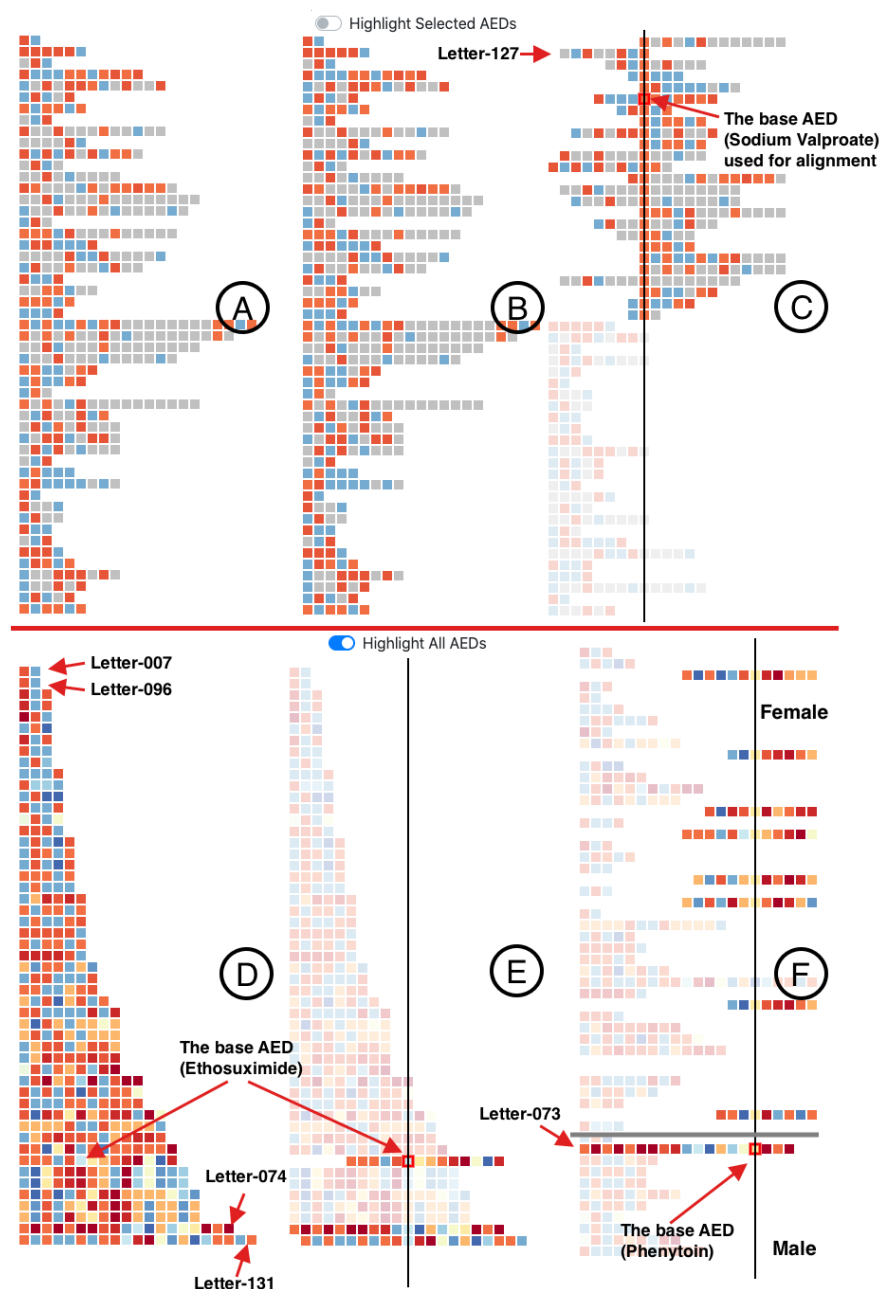


Figure 3.4: Illustrations of letter alignment in the drug chain view. A) The initial layout of the drug chain view. B) Letters are aligned via clicking on a base AED, Lamotrigine, highlighted with a red border in the first letter. Letters without the AED are shown in context with reduced opacity. C) Letters are sorted by alignment, with context letters being shifted to the end of the queue. Hovering over any block will display a tooltip containing the letter title and AED name. D) All AEDs are put in focus mode (in color) via a toggle. Chains are then sorted by the number of AEDs. E) Ethosuximide, a rare prescription in the data set, is selected as the base AED for alignment. F) Chains are sorted by gender, with a horizontal grey bar as the separator. The same subset of letters is used in this figure.

context. Chains can be interactively aligned between letters and sorted as shown in Figure 3.4 (T5). Unusual chains immediately stand out. In Figure 3.4B, a user selects

Example Query	Result
epilepsy	letters contain 'epilepsy'
epilepsy <i>AND</i> seizure	letters contain both 'epilepsy' and 'seizure'
epilepsy <i>OR</i> seizure	letters contain either 'epilepsy' or 'seizure'
epilepsy <i>NOT</i> seizure	letters contain 'epilepsy' but no 'seizure'
(epilepsy <i>OR</i> seizure) <i>AND</i> DRUG	letters contain either 'epilepsy' or 'seizure', and all AEDs from the drug list
epilepsy <i>AND</i> seizure <i>NOT</i> DOB	letters contain both 'epilepsy' and 'seizure', but no 'Date of Birth'

Table 3.2: Boolean operators *AND*, *OR* and *NOT* are supported with ( and ) for grouping. Categories are in all capitals and can be used in the query to highlight all items belonging to the category. The colour legend in Figure 3.3A left shows the list of available categories.

Lamotrigine to specify it as the base AED. All remaining chains are then aligned by the first appearance of the base AED. Chains with no matching AED are rendered as context. See Figure 3.4E.

## Advanced Visual Queries

LetterVis supports a subset of Apache Lucene [149] syntax, namely boolean operators and grouping. The incorporated boolean operators provide users with the flexibility to include or exclude keywords from the result. The colour legend is then updated to include random colours assigned to each keyword in the query history, an example is shown in Figure 3.7. The implementation aims to provide the visual means for the user to query and filter letters (T4). A query can be constructed in multiple ways:

- Clicking on cells in the matrix view will execute a query formed as 'AED on the y-axis *AND* AED on the x-axis'
- The user can drag categories from the colour legend to the search bar
- Under 'Click and Search' mode, the user can select any centroid from the global view to populate the search bar

We store a list of user-specified search queries to support undo and redo functions (T6. See Figure 3.3A right). The user can use the checkbox located in front of each query to toggle the visibility of the corresponding query and its results.

## Rendering Options

We provide three rendering options for all three visual elements (T5), *centroid*, *individual sample* and *edge* in both the global and thumbnail views. *Focus* shows the data in colour (see the user-chosen centroid and its individual samples in Figure 3.5), *Context*

shows the data in greyscale (see context centroids and individual samples in [Figure 3.5](#)) and *Hide* removes the data samples and edges. In the drug chain view ([Section 3.5.3](#)), an option is provided to highlight the data in context.

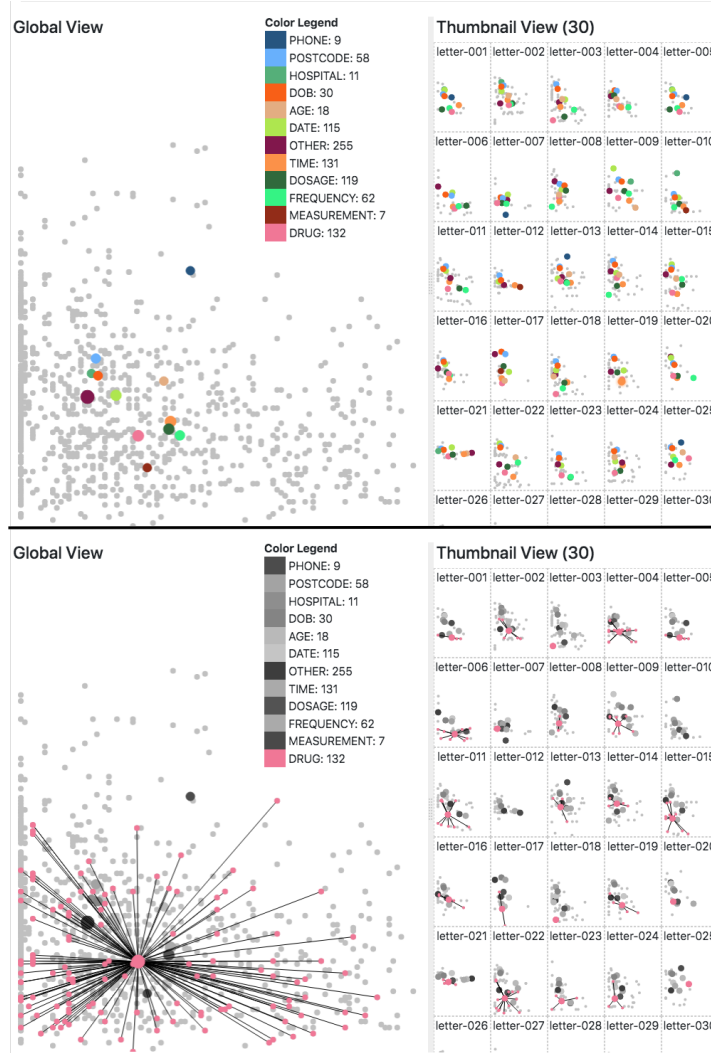


Figure 3.5: An illustration of centroid exploration in the global and thumbnail views. The top shows the default search and rendering when loading 30 letters. By default, the global view searches and renders 12 text data categories in focus and renders individual data samples in greyscale, a classical focus+context approach. A thumbnail view is presented with identical default rendering options for individual letters. The bottom shows the edges connecting the user-chosen (focus) centroid, *DRUG*, and individual samples in both views. Other centroids and individual samples are rendered as context. Edges can also be hidden as an option.

## Sorting Options

We provide 11 options to sort individual letters in the thumbnail view and the abstraction of AEDs in the drug chain view (**T5**). Sorting letters by a user-chosen dimension

gives users the control they need to find patterns and outliers quickly. For example, outliers will stand out with long edges in the global view and thumbnail views. In addition, cells in the matrix view can be sorted alphabetically or by co-occurrence. We demonstrate the benefit of having a wide selection of sorting options in our case studies in [Section 3.6](#).

## 3.6 Evaluation

Our evaluation comprises of three case studies, described in detail in [Section 3.6.1](#) and feedback from domain experts in [Section 3.6.2](#). [Section 3.6.3](#) includes reviews written by two health data analysts.

### 3.6.1 Case Studies

Each case study is motivated based on the discussions with domain experts in EHR analysis. The first case study aims to identify commonly and rarely co-prescribed AED combinations. The second case study tests the ability to find patient outliers. The third case study explores the relationship between pregnancy and AEDs. All case studies are based on 200 anonymised clinic letters described in [Section 3.3](#).

#### Case Study 1 - Identifying Common and Unusual AED Co-prescriptions

Clinicians are generally confident in prescribing the most suitable first drug, however, co-prescribing a second drug is always challenging. Visualising the common and unusual co-prescriptions may help the clinician with the decision and potentially reduce unnecessary co-prescription trials needed on patients.

After loading all letters, a matrix view is automatically generated with 18 AEDs, shown in [Figure 3.3B](#). The matrix by default is sorted alphabetically by AED co-occurrence. We then sort the matrix view by frequency on both axes from top-left to bottom-right. We immediately observe the top two AED co-occurrences near the cluster at the upper-left corner, Levetiracetam-Lamotrigine ([Figure 3.3B<sup>1</sup>](#), 45 co-occurrences) and Levetiracetam-Sodium Valproate ([Figure 3.3B<sup>2</sup>](#), 37 co-occurrences). We select these two pairs of AEDs and obtain 51 letters. This effectively filters the data with the

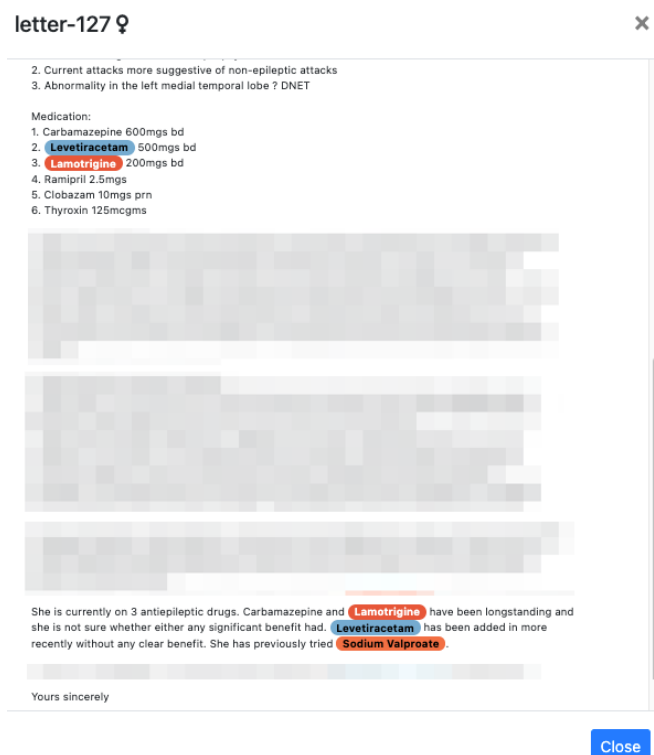


Figure 3.6: A screenshot of one letter’s original view, with personally identifiable information redacted. Search query terms are highlighted in their corresponding colours, as described in [Section 3.5.2](#).

following query: (Levetiracetam **AND** Lamotrigine) **OR** (Levetiracetam **AND** Sodium Valproate). [Figure 3.3D](#) and [E](#) show the corresponding AEDs in the global and thumbnail views. Centroids representing each AED are connected by an edge if they are connected by an **AND** operator in the query.

We then sort the letters by total edge length, the sum of distances between edges in each letter. In the thumbnail view, we observe that the centroids are more sparsely placed in letters such as letter-022 and 054 than their peers. According to domain experts, this often indicates that AEDs appearing in them are not co-prescriptions but previous medications or new recommendations. Centroids in letters such as letter-073, 133 and 174, are closely co-located, which often represents co-prescriptions. We manually inspect these letters in the focus view. The finding confirms these hypotheses.

## Case Study 2 - 5 Different Ways to Find Outliers

In this case study, we examine LetterVis’ ability to identify letter outliers. A letter with abnormal patterns often carries valuable information that requires the analysts’

attention. An outlier may also indicate an error.

In [Figure 3.4B](#), we select Sodium Valproate, due to its popularity in co-prescriptions, in the first letter to align the drug chain view. This sets it as the base AED for alignment. All remaining letters are then aligned by the first appearance of the base AED. Letters without the AED are rendered as context. See [Figure 3.4B](#) and [Figure 3.4E](#). We then sort the letters by alignment for further filtering in [Figure 3.4C](#). (1) We are able to identify one outlier immediately, letter-127, as it is the only chain ending with Sodium Valproate. We inspect letter-127 in its focus view ([Figure 3.3F](#)) which indicates Sodium Valproate was recently prescribed. According to E1, *“this is abnormal as Sodium Valproate is ususally the first or second AED for epilepsy patients.”* We expand the focus view to show the entire letter ([Figure 3.6](#)). We discover that multiple AEDs have been prescribed to the female patient with no clear benefit.

We then sort the letters by the number of AEDs ([Figure 3.4D](#)) in them and explore letters on both sides of the sort spectrum. (2) Letter-007 and 096 mention only two AEDs, we observe that the patients in the aforementioned letters are suspected cases awaiting further diagnosis. Whereas letter-074 (19 AEDs) and 131 (21 AEDs) represent confirmed patients with a history of epilepsy of more than 15 years.

Ethosuximide, shown in light blue in [Figure 3.4D](#), has significantly fewer appearances than others. (3) We align the drug chain view by Ethosuximide ([Figure 3.4E](#)), and discover that it is only co-prescribed to three patients (letter-060, 074, and 131) that are on Levetiracetam with Lamotrigine or Sodium Valproate. We view these three letters in the focus view and identify vomiting as an adverse effect caused by Ethosuximide for the patient in letter-074. The other two patients experience no benefit from Ethosuximide.

When chains are sorted by gender ([Figure 3.4F](#)), (4) we find letter-073, the only male patient (24 years old with over 23 years of history of epilepsy) in the cohort that has been prescribed Phenytoin. The patient is also the only male to receive Lacosamide.

Individual samples in the global view can also be used for finding outliers.(5) We explore two outlier patients (see red arrow in [Figure 3.3D](#)). While both patients prescribe Lamotrigine, the patient in letter-186 (top-left corner) is tapering it off as it has no clear benefit in containing seizures. In contrast, the patient in letter-104 (top-right

corner) is building up the dose as a replacement for Carbamazepine.

### Case Study 3 - AEDs and Pregnancy

In this case study, we evaluated LetterVis’ ability to identify and explore AEDs prescribed to patients with planned or ongoing pregnancies. The prevention of adverse effects of AED is an important research topic. By using LetterVis’ advanced visual interface, the user can combine any keywords to study their associations and reveal insightful patterns. The case study is inspired by research on the effects of AED in pregnancy [212], the research indicates that in-utero exposure to Sodium Valproate is likely to negatively affect a child’s cognitive ability, while Lamotrigine and Carbamazepine have little or no impact.

We first construct the query ‘(pregnant *OR* pregnancy) *AND* DRUG’ to filter out 182 letters. The query highlights both *pregnant* and *pregnancy* with all 50 AED names supplied by our domain expert partners. We then apply sorting by gender to obtain the thumbnail view shown in Figure 3.7. Two letters include two male patients describing their own births, we exclude these two. We classify the resulting 16 letters into five categories as shown in Table 3.3. All patients in categories 1, 2, and 5 were informed of the teratogenicity (the property or capability of producing congenital malformations.) of AEDs which may result in birth defects.

Letter-011 indicates a high-risk case, a 52-year-old patient who suffered four seizures in eight months is planning pregnancy. In this special case, a higher than usual dose of Folic Acid is prescribed to help prevent birth defects. In letter-060, the physician proposed multiple AED co-prescriptions to gradually replace Sodium Valproate, in order to prepare the patient for pregnancy. In letter-144, Sodium Valproate is showing a remarkable effect in reducing frequency of seizures for the patient. Because she did not have a planned pregnancy, the physician decided to increase the dose.

Letter-093 contains a special case where a female is suspected to suffer from nonepileptic psychogenic seizures, hence common AEDs such as Levetiracetam, Lamotrigine, and Sodium Valproate were never prescribed.

During the process, we also discover two identical letters (letter-103 and 107) using different pseudo names, this is likely due to human error during the manual anonymi-



Figure 3.7: We execute the query ‘(pregnant *OR* pregnancy) *AND* DRUG’ and sort letters by gender to focus on pregnancy. The thumbnail view shows two male patients separated from the rest by a grey bar. High-risk cases discovered in Case Study 3 are highlighted with a red line.

sation process. Please see the accompanying video for a demonstration of these case studies.

### 3.6.2 Domain Expert Feedback

We regularly demonstrate LetterVis to our domain expert partners (E1 and E2) to guide the development and present intermediate results. We also provide a live version available online for them to explore. We provide excerpts of their feedback below. In general, our collaboration process adheres to the Visual Information-Seeking Mantra [9]: 1) we demonstrated the global view that shows an *overview* of the data to experts,



1. Pregnancy Planned	2. No Pregnancy Planned	3. Pregnancy Completed	4. Irrelevant Case	5. Special Case
letter-11	letter-20	letter-72	letter-115	letter-93
letter-60	letter-21	letter-87	letter-131	
letter-81	letter-144	letter-127		
letter-103	letter-151			
letter-107	letter-176			
<b>Total:</b>	5	5	3	2

Table 3.3: The resulting letters after searching and filtering based on pregnancy and AED. Letters are classified into five categories: 1) Patients with pregnancy planned or potentially planned. 2) Patients with no pregnancy planned. 3) Patients with pregnancy completed. 4) Letters containing keywords ‘pregnancy’ and ‘pregnant’ but are irrelevant to the patient’s condition, such as describing the patient’s own birth. 5) A special case is described in detail in Case Study 3 - AEDs and Pregnancy. **Green** highlights letters with Sodium Valproate prescribed.

2) they then requested to *zoom and filter* the outliers found in the global view. This is fulfilled by the thumbnail view, 3) eventually *details were demanded* to verify any findings from previous stages, through both the focus view and chain view.

During our first demonstration of LetterVis with E2, we demonstrated the coordination between the global, thumbnail, and focus views, the expert immediately commented, *“That’s interesting, if you can spot pregnancy and certain drugs are in close proximity, you immediately want to read that letter because something is not right. I can see it’s been really useful”* (R2). E2 pointed out that the limitation of our simple numerical approach is that the user is unable to query for words. The expert was also particularly interested in seeing a view that’s specifically tailored for visualising AEDs, with the ability to use syntax-based search queries to improve the exploration of items of interest (R4). We implement his recommendation, Apache Lucene, into our next version (R7, T5).

In our second feedback meeting with E1 and E2, immediately after the feature introduction, both experts were able to picture a use case for identifying outliers by using deviated centroids in the global and thumbnail views, E2 stated that, *“If you are able to see centroids representing AEDs in one letter deviated far more than the global trend, we don’t trust this letter, we might need to investigate the prescriptions in that particular letter”* (R2). Both experts were keen on visualising AED co-prescriptions, and the demonstration spent 30 minutes discussing this topic. E1 pointed out that, *“One useful use case I can imagine is to visualise combinations of AEDs for different patients and even how often one AED is mentioned together with another. It’s really hard to conduct clinical trials for more than one AED, clinicians usually know what the*

*best first AED is, but not the second. Designing a trial for that is nearly impossible” (R8). E2 further elaborated that, “Some patients on multiple AEDs might have multiple seizures per week, but when they are given only one AED, their seizure frequency might be reduced to one per week. One existing tool we are using still relies on the command line to operate, so this dashboard-like visualisation tool can be really helpful”.*

We introduce the matrix (Figure 3.3B) and drug chain view (Figure 3.3C and Figure 3.4) for visualising AED co-prescriptions and prescription evolutions (T2). We demonstrated these two visual designs to E1 and E2 during our third and fourth feedback sessions. E2 commented that, *“This is definitely useful, the way we are currently doing is really laborious, we have to go through the patient’s EHRs and sequentially look at what AEDs were given at each visit. The drug chain view is looking at the problem we need to solve” (R8).* When the drug chain view was sorted by the number of AEDs and aligned the chains by a rare AED, E1 was immediately able to identify an outlier patient that is on an unusual AED co-prescription, *“Pregabalin and Retigabine are a very strange combination, I never thought of searching for that, I’m definitely going to look at that patient”.*

We demonstrate LetterVis and the case studies in a supplementary video: <https://youtu.be/jSVzhCjLi-U>.

### 3.6.3 Domain Expert Review

The following written feedback was provided directly by the experts (E1 and E2).

“LetterVis presents a novel way to visualise trends and potential outliers across sets of clinic letters. Unstructured texts have not traditionally been used as a data source for analysis in healthcare, as the data is not available in a readily parseable format such as structured data. Recent advances in the field of Natural Language Processing have yielded NLP methods to extract structured data from clinical prose [153], where more traditional analyses can take place. LetterVis employs NLP and data visualisation techniques to help isolate and communicate important trends to the user.

“Many clinical decisions can be improved by the analyses that LetterVis offers. For example, there is limited evidence on the best antiepileptic drug (AED) combination to use for patients with severe epilepsy [139]. The co-occurrence matrix can visualise AED

combinations across a range of frequencies to potentially identify the most common and/or stable AED combinations. Less frequently used combinations can also be easily identified. A limitation of the co-occurrence matrix is addressed in the drug chain view – namely the ability to visualise patients who are prescribed more than two AEDs, and it is immediately clear which patients may be problematic based on the number of different AEDs appearing in the chain.

“By loading AEDs vs side effects into the co-occurrence matrix it is possible to view associations between AED and side effects. The thumbnail view adds more granularity to this analysis by presenting the proximity of side effects to the AED in question. In 2018 the Medicines & Healthcare products Regulation Agency strengthened their position on the avoidance of valproate to be used in women and girls, especially during pregnancy [214]. Therefore, it is important to determine the likelihood that letters that mention pregnancy and valproate fall into categories around the education of valproate in pregnancy, or if valproate is still used during pregnancy. The thumbnail view is an ideal tool to hone in on letters that may show the latter case because currently prescribed AEDs are usually found at the beginning of the letter, where pregnancies will be mentioned later in the document.

“Some future work might include the integration of medical ontologies to rapidly build code lists of interest that can be used in the already highly configurable Lucene-based query capabilities. Given that context is very important when using search terms, i.e. negation or hypothetical discussions around AED side effects, adopting more advanced NLP techniques or integrating with existing technology such as GATE [120] or cTAKES [364] would help increase user confidence in any trends presented to them. LetterVis could potentially expand to present entire timelines for individual patients and not be restricted to analyses within one letter. For example, the chain view could be used to align newly diagnosed patients and their first AED to determine popular first-line drugs, and how they change over time, or to monitor side effects and frequency of seizures when a new AED is detected in subsequent letters.

“LetterVis is well positioned to take advantage of the emergence of unstructured data being used for healthcare research, and its methods will offer clinicians the vital tools they need to see the big picture across potentially millions of clinic letters.”

## 3.7 Limitations and Future Work

Our work is currently limited by data set size, with larger data sets from an NHS Healthboard pending approval. This expansion will allow us to address scalability issues, as the current browser-based implementation depends on local computational power, which slows performance with larger data sets. A cloud-based solution could mitigate this, but privacy and security challenges inherent to EHR data pose significant hurdles.

The current rule-based NLP approach, developed with domain experts, is effective but limited. Incorporating advanced NLP techniques could improve extraction accuracy and automate processes, such as analysing AED side effects and interactions. This would reduce reliance on manual querying and enhance usability for broader clinical applications.

While the modular design of LetterVis supports epilepsy-specific research, additional modules can be easily added for other medical domains. Expanding display space for simultaneous rendering of views could also improve usability and efficiency.

Currently, the drug chain view only shows AED sequences, not co-prescriptions, which still require human verification. Future NLP improvements may help in this area, but will not eliminate the need for manual validation (**R3**, **R5**, **T3**).

We will explore the use of multiple colormaps and their effects on users with colour vision deficiencies. The choice of colormaps was heavily influenced by input from domain expert partners, who provided valuable guidance based on their professional experience and visualisation needs. However, these experts did not have colour vision deficiencies, which may have limited their awareness of accessibility issues related to colour perception. While the colormaps were selected to maximise clarity and align with the specific tasks required for clinical text analysis, this design decision inadvertently prioritised domain-specific preferences over broader accessibility considerations.

## 3.8 Conclusions

In this chapter, we present a novel visualisation tool, LetterVis, to support the analysis of clinic letters through advanced interactive visual designs and queries. The work aims

to support EHR researchers to explore free-text EHR data and address their research hypotheses in a transparent and explainable manner. The strength of this work is the novel concept of letter-space and how it is applied to a real-world problem. Through our collaboration with three domain experts, we identify and address a selection of important tasks via customised letter-space designs and interactive user options. We incorporate NLP techniques to preprocess clinic letters with a list of extraction rules curated together with our domain expert partners. We then develop an advanced visual query interface that includes five customised visual designs to support the analysis of EHR free text data. We demonstrate LetterVis with three empirical use cases inspired by real-world scenarios. In-depth evaluations are also conducted with domain experts. Its reliance on predefined data categories and the need for significant user familiarity with the tool may limit its accessibility to nonexpert users. Additionally, while the visualisations provide detailed insights, the complexity of the interface and potential cognitive overload from high-dimensional data may require further refinement to enhance usability. Future iterations could focus on simplifying interactions and integrating more automated analytical features to further broaden its applicability and impact.

LetterVis serves as a foundational example of how visualisation can transform unstructured clinical data into actionable insights, setting the stage for the broader themes explored in the thesis. Its focus on clinical letters highlights the challenges of working with textual data in EHRs, such as extracting meaning from fragmented, context-rich narratives. This chapter directly connects to later chapters by addressing the broader problem of representing heterogeneous and complex healthcare data. For instance, the principles of abstraction and interactivity in LetterVis are also leveraged in the geospatial and time series visualisations introduced in subsequent chapters. By tackling text-based data early on, the thesis builds a cohesive narrative that demonstrates how visualisation methods can adapt to diverse EHR challenges while maintaining a user-centred approach throughout.

## 3.9 Appendix

### 3.9.1 List of Expert Sessions

Session	Date
Expert 3 Initial Interview (in person)	7 June 2019
Expert 1 & 2 Initial Interview (in person)	29 June 2019
Expert 2 Feedback (virtual)	7 April 2020
Expert 1 & 2 Feedback (virtual)	9 July 2020

Table 3.4: Summary of domain expert sessions (both interviews and feedback), conducted in person and virtually due to the COVID-19 pandemic restrictions.

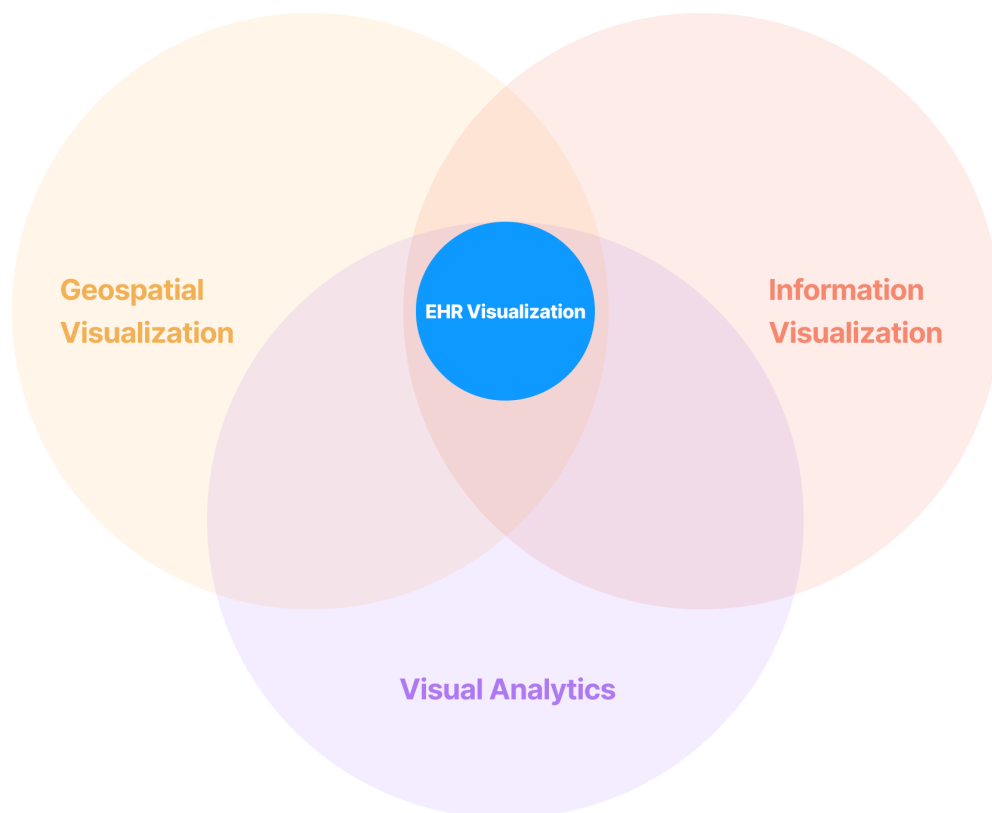
### 3.9.2 List of Interview Questions

1. What is your background, including education and previous occupations?
2. Please provide a brief description of your current occupation.
3. What area does your research cover?
4. What are your research objectives?
5. What are your sources of data?
6. Do you have any hypotheses?
7. What general types of data set are common in your research (e.g. numerical data, text documents, images, etc.)?
8. On average, what is the size of data set do you generally work with? (e.g. 5,000 rows of CSV, 3,000 images, etc.)
9. What methods do you use to analyse the data (e.g. statistical, excel, document analysis tools, etc.)?
10. What methods or tools do you use to get the “bigger picture” (or an overview) of your collections of data?
11. What are the areas of improvement you’d like to see in the tools you use for your research(s) (e.g. data processing speed, interactive functions, user-friendliness etc.)?
12. What do you consider to be your biggest challenge in terms of answering research questions?
13. If you could find out anything in your research(s), what would that be?
14. What are your unsolved research problems?

## Chapter 4

### Demers Cartogram with Rivers

Wang, Q., Xu, K., & Laramee, R. S. (2024). Demers cartogram with rivers. Visual Informatics. <https://doi.org/10.1016/j.visinf.2024.09.003>



*“The single biggest threat to the credibility of a presentation is cherry-picked data.”*

– Edward R. Tufte, the Father of Information Design (1942 - present)

The chapter is based on the publication *Demers Cartogram with Rivers* in *Visual Informatics* [319]. By extending the Demers Cartogram, we introduce a hybrid cartogram layout algorithm that incorporates dynamic topological features, such as rivers, to improve legibility, readability, and overall accuracy of the information presented in Demers Cartogram. The idea is inspired by previous research [202, 218] identified in our survey in Chapter 2.

## Contents

---

<b>4.1</b>	<b>Introduction</b>	<b>103</b>
<b>4.2</b>	<b>Related Work</b>	<b>104</b>
<b>4.3</b>	<b>Data Description</b>	<b>108</b>
4.3.1	Choropleth Shapefile	108
4.3.2	River Shapefiles	109
4.3.3	EHR Data	110
<b>4.4</b>	<b>Demers Cartogram with Rivers</b>	<b>110</b>
4.4.1	Initialisation with Rivers	111
4.4.2	Node Layout and Overlap Removal	111
4.4.3	River Intersection Testing	113
4.4.4	Translating Rivers	114
4.4.5	Process Stalemates	116
4.4.6	Terminating the Algorithm	118
4.4.7	User Options	119
<b>4.5</b>	<b>User Evaluation</b>	<b>120</b>
4.5.1	Study Hypothesis	121
4.5.2	User Study Variables	121
4.5.3	User Study Design	122
4.5.4	User Study Analysis	125
<b>4.6</b>	<b>Limitations and Future Work</b>	<b>130</b>
4.6.1	Experimental Design Confounds	130
4.6.2	Colormap choice	131
4.6.3	Overlap removal algorithm choice	131
4.6.4	Generalisability	132
4.6.5	Improved User Study	132
<b>4.7</b>	<b>Conclusions</b>	<b>132</b>
<b>4.8</b>	<b>Appendix</b>	<b>133</b>
4.8.1	Preprocessing Shapefiles	133
4.8.2	List of Likert Scale Questions	139

---



## 4.1 Introduction

Cartograms are representations of geographical and abstract data based on a value-by-area mapping combining statistical and geographical information [66, 94]. Various styles of cartograms have been proposed and implemented for applications such as urban planning [211, 287], natural hazard forecasting [243, 269], conservation and environmental planning [208, 244], political and social demographics [205, 284], and decision-making for public health [259, 294].

Among the four types of cartogram categorised in a survey by Nusrat and Kobourov (See Section 4.2 for definitions of contiguous, noncontiguous, rectangular, and Dorling), a trade-off is made between types of accuracy (See Table 4.2). For this project, we focus on noncontiguous cartograms like Demers cartograms because they facilitate statistical comparison between regions, they can make good use of screen space, and comparison of regions is useful when studying Electronic Health Records (EHR) data. Demers cartograms offer the advantage in cases where the data is not directly correlated to region sizes. In other words, Demers cartograms are useful when a data dimension is not describing the geography of the region it represents but is tied to something else, for example, the health of its population. In addition, the comparison of magnitudes becomes an area estimation task, which is effective for numeric data encoding [138]. See Nickel et al. for a more complete description of the advantages that Demers cartograms offer. One of the drawbacks of Demers cartograms is that they may become more difficult to read when a region becomes displaced from its geospatial origin through the node layout process. The layout of these more abstract shapes may simultaneously reduce the map’s legibility and increase error. See Tong et al. for a more detailed explanation. Building on Demers cartograms [25], we introduce and develop novel features, such as rivers, with the aim of improving readability and geographical accuracy without sacrificing statistical accuracy. Standard Demers cartograms are composed of square nodes that represent geographic enumeration units. As such, this can reduce their legibility. We implement a new hybrid cartographic layout algorithm that combines rivers with the placement of nodes representing geographic enumeration units, in this case a Clinical Commissioning Group (CCG). We hypothesise that the introduction of rivers improves the overall legibility of a cartogram. By *legibility* we mean readability

and the ability to interpret the cartogram. To assess this hypothesis, we designed an experimental setup where participants engaged in correspondence and location tasks as part of a user study. To reduce error and make efficient use of screen space, the algorithm also updates the position of rivers to accommodate the node layout. We then apply the algorithm to a real-world case study using EHR data to evaluate the result. We present a user study that demonstrates its effectiveness.

Our contributions include:

- A new variant of Demers cartograms that incorporates rivers to improve readability and recognisability, as shown in [Figure 4.7](#)
- A novel hybrid layout algorithm that combines node positions with features such as rivers, as described in [Figure 4.2](#) and [Algorithm 1](#)
- A user study evaluation ([Section 4.5](#)) of the technique with an application to EHRs.

The results of the user study indicate that rivers can improve the legibility of cartograms. One of the major challenges involved is developing a layout algorithm that handles different shapes. In other words, the hybrid layout algorithm is novel because it handles different types of elements: square node representing regions and polylines representing rivers. Another challenge we overcome in developing the algorithm is to resolve stalemate situations where nodes become congested due to constraints imposed by rivers, while ensuring error minimisation.

## 4.2 Related Work

This section introduces the characteristics of various cartogram styles, describes relevant applications of cartograms, and provides a brief overview of some real-world implementations of cartogram-based visualisations.

### Definitions:

While we focus on rectangular cartogram variants, we start with brief definitions of contiguous and noncontiguous cartograms: Contiguous cartograms preserve topology, maintain connectivity with their adjacent neighbours, but are also subject to distortion

Literature	Cartogram Type(s)	Geographic Region(s)	Number of Nodes	Year
Warf and Winsberg	Dorling	US	3,142	2008
Sun and Li	Dorling, Mosaic, neighbour-preserving	US, China	34 - 49	2010
Cruz	Dorling, Noncontiguous, neighbour-preserving	Portugal	2,882	2017
Tong et al.	Demers	England	209	2018
Gao et al.	Dorling	China	34	2020
Nusrat et al.	Contiguous, Dorling	US	49	2020
Nickel et al.	Noncontiguous, Demers	US, Netherlands, World	49 - 342	2022

Table 4.1: Related work with noncontiguous cartogram-based visualisations. **Cartogram type** is the type of cartogram used. **Geographic region** is the geographic region depicted by the cartogram. **Number of nodes** is the number of nodes (representing geographic enumeration units) depicted in the cartogram.

in shape. Noncontiguous cartograms sacrifice topological connectivity with neighbours to enable expansion or reduction in size while maintaining their polygonal shape [66].

Nusrat and Kobourov define and summarise three major accuracy dimensions for cartograms: statistical, geographical, and topological. Each cartogram design may make various types of accuracy trade-offs between dimensions. We provide a comparison of these trade-offs as introduced by Nusrat and Kobourov in Table 4.2. In addition, we include the Demers as it is the focus of our work.

Dorling cartograms, as a variant of noncontiguous cartograms, generally do not preserve geography and topology. A Dorling cartogram is statistically accurate, regions are represented by circles and the data dimensions of interest are represented by the circle area [87]. In a Demers cartogram, a variant of Dorling, squares are used instead to capture a certain level of topology, as described by Cano et al. in their related work section. Dorling cartograms are unable to maintain topological accuracy as circles are often repositioned to remove overlap. Here we focus on Demers cartograms as we use squares to depict regions. This style of cartogram offers the advantages that the comparisons between regions are intuitive and screen space utilisation is more efficient. This is important in our use case scenario involving EHRs. Demers cartograms, where regions are represented by squares, often have the advantage of preserving a higher level of topology at the cost of geographical accuracy [25].

Rectangular cartograms are contiguous and do not preserve geographical accuracy [3, 36]. Depending on the variant, a rectangular cartogram may trade-off between statistical and topological accuracy.

Mosaic cartograms usually use square or hexagonal tiles to depict regions, and

are contiguous and sacrifice statistical accuracy to preserve some level of geographical accuracy [146]. Some variants are able to preserve topological accuracy as well.

### Peer-reviewed Applications:

There has been a substantial amount of research done in this area, here we review some of the important work that has inspired our work. Warf and Winsberg [62] use a Dorling cartogram to represent religious diversity in the US. Sun and Li [82] depict 1996 US election data and 2005 Chinese population data using Dorling, Mosaic, and contiguous cartograms. Gao et al. [259] present a Dorling cartogram to illustrate COVID-19 infections in China. Tong et al. [218] use a Demers cartogram to visualise health-related data by CCG regions in England, the work introduces a novel technique to remove the overlap of squares based on topological features, aiming to improve both geographical and topological accuracy. Nusrat et al. [268] investigate the memorability of contiguous and Dorling cartograms using multiple data sets that include demographic, agriculture, and retail data in the US. See Table 4.1 for a list of literature that adopts cartograms for visualisation with corresponding geographical regions and node counts.

Our work extends the algorithm described by Tong et al. which incorporates a static topological feature into Demers cartograms. Our work enhances that of Tong et al. in multiple ways. First, we introduce multiple features (rivers) into the layout, as opposed to a single river. Second, we make topological features dynamic and further improve legibility and geographical accuracy. By the term *dynamic*, we mean that the position of the rivers is updated as part of the layout algorithm. In previous work, the river is static and serves merely as a boundary. To illustrate this more clearly, we refer to

Cartogram Variant	Accuracy			
	Statistical	Geographical	Topological	Contiguity
Contiguous	Variable	Variable	Accurate	Yes
Noncontiguous	Accurate	Shape is accurate	Inaccurate	No
Rectangular	Variable	Shape is inaccurate	Variable	Yes
Dorling	Accurate	Inaccurate	Variable	No
Demers	Accurate	Inaccurate	Variable	No

Table 4.2: ■ Trade-off between dimensions. ■ Dimension sacrificed in order to improve ■ target dimension’s accuracy.

the video demonstration published by Tong et al. Because the behaviour of the layout algorithm is dynamic, a video is more appropriate to convey the motivation of dynamic features of interest. We can observe the cartographic layout algorithm at 1:30 and 2:40 of the video by Tong et al. During the update process, we can clearly observe nodes cross the River Thames as the size expands. As the size of all the nodes expands, some nodes are pushed south to make more use of screen space. If we want to make the most efficient use of screen space, we need to translate the River Thames further south during the layout process to prevent nodes from crossing it. This is one of the main motivations for introducing dynamic rivers. By introducing river translations, our implementation prevents nodes from crossing rivers, thereby improving screen space efficiency and cartogram legibility. For comparison, see our video demonstration at <https://youtu.be/PRNEF3J1hl0> (from 0:28 to 0:32 and from 1:49 to 1:56).

Third, we improve the algorithm to resolve stalemates. Finally, the way we evaluate the cartograms is also different. Tong et al. count river crossings to evaluate error (a statistical metric). Here our focus is on readability, and thus we include a user study.

### **Cartograms in Media:**

Cartograms are an engaging visual representation and therefore they are a popular choice of representation in covering various topics by the media. The Washington Post uses cartograms to visualise the US overseas economic assistance, in arms sales (Mosaic) [158], the 2016 US Election (contiguous) [162], and the Brexit Referendum (Mosaic) [181]. National Geographic uses contiguous and Mosaic cartograms to analyse the 2016 US Election results [174], the same topic is also covered by the Financial Times with a Dorling cartogram [180]. Cruz adapts a Dorling cartogram with both contiguous and noncontiguous cartograms to represent the gender pay gap in Portugal. Sandberg reports the 2018 US midterm Election with a Mosaic cartogram, the same approach is used to cover the 2020 US Election by the New York Times [276] and Bloomberg [266].

One of the disadvantages of Dorling and Demers cartograms is legibility. The layout algorithms may displace regions far from their original position and make the maps more difficult to interpret. Nickel et al. present a method to compute stable Demers cartograms with multiple constraints to maintain adjacencies with no overlapping nodes.

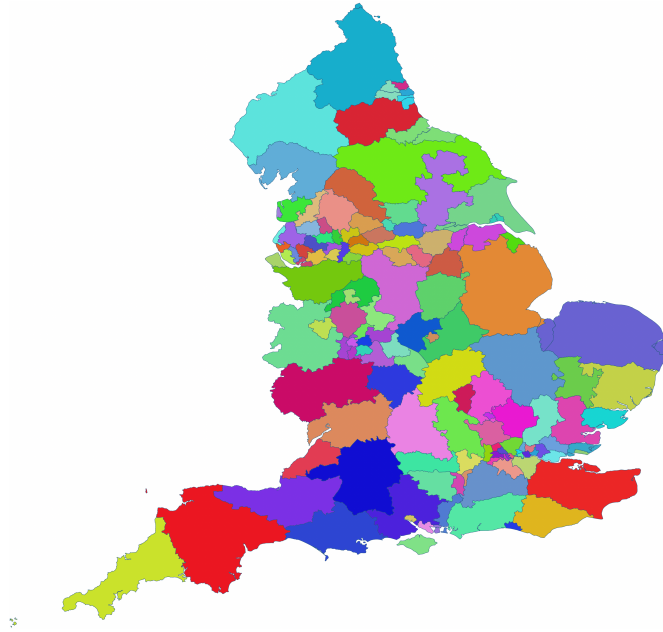


Figure 4.1: A map of 135 CCGs in England as of 2020, obtained from the Open Geography Portalx [352] with EPSG:4326 (WGS84 - World Geodetic System) as the Coordinate Reference System (CRS).

In this chapter, we introduce a new type of topological feature, a river, as a constraint to compute the final layout, with the aim of improving the interpretation, readability, and accuracy of this class of cartograms.

## 4.3 Data Description

Processing heterogeneous data can be challenging, especially when an EHR data set is involved, because the data comes from multiple sources [311]. The first step is to obtain both geospatial boundaries and EHR data. The second step is to preprocess the EHR data to remove empty and erroneous values. The final step is to transform the data into a suitable format for cartograms. Geospatial boundaries, or shapefiles, were obtained from the sources described here.

### 4.3.1 Choropleth Shapefile

Clinical Commissioning Groups (CCGs) are the primary administrative and geographic unit of the National Health Service (NHS) in the UK [350]. The number of CCGs changes over time due to NHS reorganisation. The most up-to-date shapefile is available

```

relation(2263653);>>;
// River Great Ouse: 2798097
// River Trent: 2863468
out skel;

```

Listing 4.1: The query that downloads the shapefile of River Thames from OpenStreetMap via the Overpass Turbo API.

Shapefile	Original	GeoJSON	TopoJSON
Rivers	2.0 MB (GeoJSON)	1.4 MB	-
NHS CCGs	46.6 MB (.shp, Esri vector shapefile)	140.2 MB	-
Merged	-	-	16.3 MB

Table 4.3: The file size is reduced by 88.5% from the original size.

from the Open Geography Portalx [352]. We decided to use the CCG shapefile from 2020 at the time of writing, due to the absence of published public EHR data based on the latest CCG reorganisations that took place in 2021 and 2022.

### 4.3.2 River Shapefiles

We used OpenStreetMap [353] as our data source to obtain shapefiles for the River Thames, the Trent River, and the Great Ouse River in England. These rivers were chosen as they are well-known rivers, pass through regions with dense populations, and provide informative geographical and topological cues. Although including smaller rivers is technically feasible, it may not increase the legibility of the cartogram. This is an open question for future work.

We first obtain a relation ID by searching for a river, e.g. River Thames, on OpenStreetMap. The relation ID is used to construct a query (see Listing 4.1) that enables the user to download the entire river shapefile using Overpass Turbo [354].

After acquiring the shapefiles, we used QGIS [355] to manually adjust the projections and convert them into GeoJSON files. Finally, mapshaper [324] is used to merge and convert the GeoJSON files into a TopoJSON file [365]. TopoJSON eliminates redundant coordinates in the data, improving the rendering speed of our implementation. See Table 4.3 for the preprocessing result.

We describe the one-time preprocessing steps in more detail in Section 4.8.1.

### 4.3.3 EHR Data

We obtained the Clinical Commissioning Group Outcomes Indicator Set (CCG OIS) from NHS Digital [351]. The OIS is a set of indicators that are used to measure the quality of care and the associated health outcomes in the NHS. Each CCG has a unique ONS code, which is used to link the CCG shapefile with the statistical data.

- Under 75 mortality: cardiovascular disease, respiratory disease, liver disease, and cancer
- Emergency hospital admission: stroke, alcohol-specific admission and readmission, coronary heart disease, re-admissions within 30 days of discharge, children with lower respiratory tract infections

For all data sets, a spreadsheet including the following is provided:

- Reporting period: Calendar year of registration
- Period of coverage: Start and end date or reporting period
- Breakdown: Organisation type
- ONS code: UK Office for National Statistics CCG code
- Level: CCG Code
- Level description: CCG Name
- Gender
- Indicator value: Directly standardised mortality rate
- CI lower: lower 95% confidence interval
- CI upper: upper 95% confidence interval
- Denominator: The count of registered patients
- Numerator: Number of deaths

## 4.4 Demers Cartogram with Rivers

[Algorithm 1](#) and [Figure 4.2](#) provide an overview of the hybrid cartogram layout process that includes rivers.



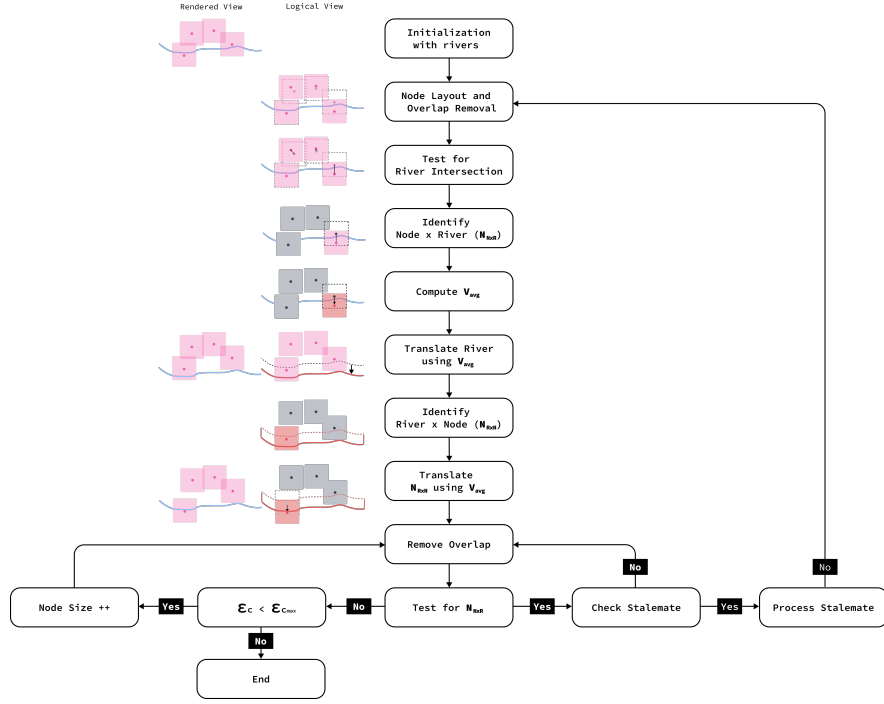


Figure 4.2: An overview of our hybrid layout algorithm incorporating rivers. See also Algorithm 1 in Section 4.4 for more detail. See Figure 4.4 for the logic of processing a stalemate. For illustration purposes, we show the rendered views alongside the logical views representing the actual computation, and use the same size for all squares for clarity.

#### 4.4.1 Initialisation with Rivers

We first load and (optionally) render the CCG geospatial boundaries. For each CCG we compute the centroid and represent it using a square node,  $n$ , with the initial size,  $s = 1$  pixel. We then load the river shapefiles and render the rivers. Since the vertices of the river in the shapefiles are not in sequential order, we first render the starting vertex, followed by the next nearest vertex. We do not need the original river resolution to incorporate them into cartograms. We reduce their resolution to match that of the cartogram nodes in order to facilitate node-river intersection tests. This rendering approach enables us to adjust the river resolution as shown in Figure 4.3. We further apply simplification by removing vertices that are too close to each other. The initialisation procedure is a one-time process that can be saved for reuse.

#### 4.4.2 Node Layout and Overlap Removal

We first apply the Fast Node Overlap Removal (FNOR) algorithm that solves the Variable Placement with Separation Constraints (VPSC) problem [44] in order to remove

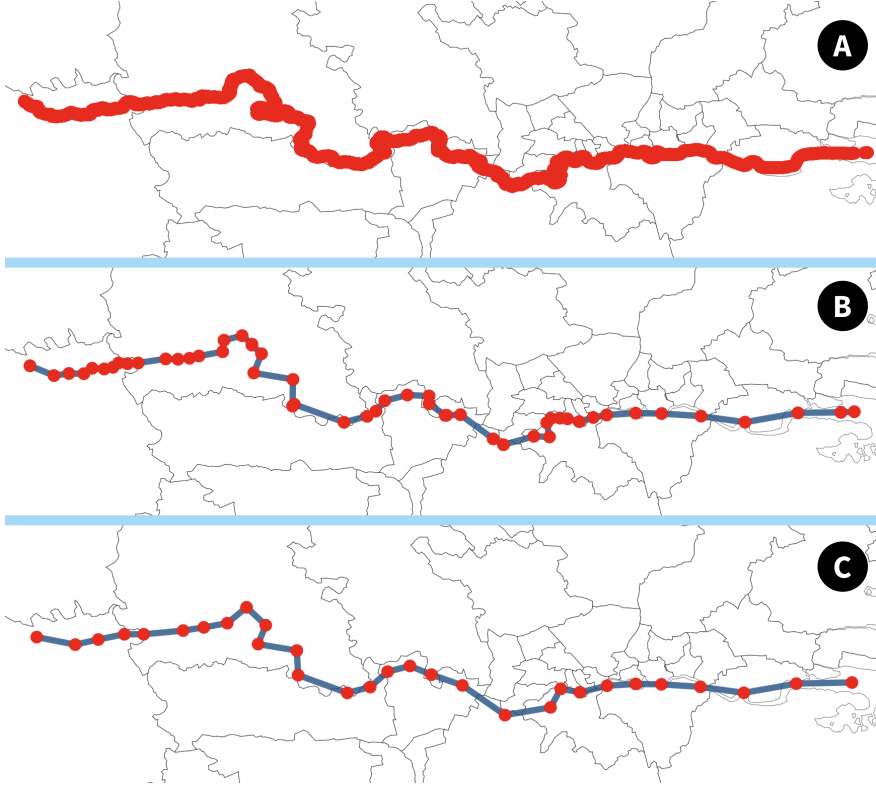


Figure 4.3: The resolution of rivers can be dynamically adjusted by the user. (A) shows River Thames at its original resolution with 10,170 edges. (B) shows the river at a reduced resolution of 49 edges. We further smooth the river by removing 19 vertices in dense areas, as shown in (C). The reduced resolution preserves the majority of River Thames’ original shape and improves the performance of our river intersection tests.

the overlap between square nodes. We chose FNOR over other node overlap removal algorithms because FNOR is able to minimise spread and movement of nodes while maintaining a good level of global shape preservation [252]. Initialise with a pixel size of unity, we gradually increase the node size by one unit at a time to ensure smooth transitions. An increase in  $s$  can cause the nodes to overlap. During overlap removal, we compute node trajectories (See Algorithm 2) and translate nodes to their new position. Nodes that cross a river, denoted  $\mathbf{n}_{n_{xr}}$ , are translated back to their previous position. If a node oscillates across a river, we identify this as a stalemate. One iteration of the layout ends when 1) no node overlap is present; AND 2) no nodes cross a river. We then increase  $s$  by one unit and repeat the algorithm until the average cartographic error,  $\epsilon_c$ , a measure set by the user, reaches its maximum value  $\epsilon_{c_{max}}$ . The gradual size increase process provides stability to the layout, as can be seen in the accompanying video.

---

**Algorithm 1** Procedure to adjust river positions, remove node overlap and prevent nodes from crossing rivers.

---

**Global variables:**

$\mathbf{L} \leftarrow$  a list of  $\mathbf{n}$  representing regions with the following properties:

$\mathbf{L.cross} \leftarrow$  the number of  $\mathbf{n}$  in  $\mathbf{L}$  that crosses a river

$s \leftarrow$  the initial size of all nodes

$\epsilon_c \leftarrow$  the average cartographic error of all nodes

$\epsilon_{c_{max}} \leftarrow$  the maximum cartographic error of all nodes

$w \leftarrow$  the maximum number of iterations indicating a stalemate

**Local variables:**

$\mathbf{n} \leftarrow$  a node is an object with the following properties:

$\mathbf{n.x}, \mathbf{n.y}$ , or  $\mathbf{n}(x, y) \leftarrow$  the x and y coordinates of  $\mathbf{n}$

$\mathbf{n.cross} \leftarrow$  the number of times that  $\mathbf{n}$  crosses a river

$\mathbf{n}_p \leftarrow$  the previous position of  $\mathbf{n}$

$\mathbf{n}_t \leftarrow$  the translated position of  $\mathbf{n}$

```

1: procedure UPDATELAYOUT
2:    $s \leftarrow 1$ 
3:   while  $\epsilon_c < \epsilon_{c_{max}}$  do
4:      $\mathbf{L.cross} \leftarrow 1$  ▷ Trigger the while loop
5:     while  $\mathbf{L.cross} > 0$  do
6:        $\mathbf{L.cross} \leftarrow 0$ 
7:        $\mathbf{L} \leftarrow \text{RemoveOverlap}(\mathbf{L})$ 
8:        $\text{TRANSLATERIVER}(\mathbf{L})$ 
9:        $\mathbf{L.cross} \leftarrow \text{TRANSLATENODE}(\mathbf{L})$ 
10:    end while
11:     $s++$ 
12:  end while
13: end procedure

```

---

### 4.4.3 River Intersection Testing

The logic for translating the position of a node is detailed in [Algorithm 2](#). When a node's position changes, we test if the node's trajectory intersects any segment of a river. See [Algorithm 3](#). A bounding box intersection test can be performed between the edge defined by the node translation and the river edges to reduce the number of edge intersection tests required. Using the intersection test, we identify all nodes that cross the river as a result of the initial FNOR algorithm. We label these nodes,  $\mathbf{n}_{xr}$ .

---

**Algorithm 2** Procedure to translate node positions.

---

**Input:**

$L \leftarrow$  a list of  $n$  representing regions

**Output:**

$\epsilon_t \leftarrow$  the number of nodes crossing the river in the input

**Global variables:**

$w \leftarrow$  the maximum number of iterations indicating a stalemate

**Local variables:**

$n \leftarrow$  a node is an object with the following properties:

$n.x, n.y$ , or  $n(x, y) \leftarrow$  the x and y coordinates of  $n$

$n.cross \leftarrow$  the number of times that  $n$  crosses a river

$n_p \leftarrow$  the previous position of  $n$

$n_t \leftarrow$  the translated position of  $n$

```

1: procedure TRANSLATENODE( $L$ )
2:    $\epsilon_t \leftarrow 0$ 
3:   for each  $n \in L$  do
4:     if  $n(x, y) \neq n_t(x, y)$  then
5:        $n(x, y) \leftarrow n_t(x, y)$ 
6:       if TESTINTERSECTION( $\overrightarrow{nn_t}$ ) = True then
7:          $n.cross ++$ 
8:          $\epsilon_t ++$ 
9:         if  $n.cross < w$  then
10:            $n(x, y) \leftarrow n_p(x, y)$   $\triangleright$  Translate back to previous position
11:         else
12:           PROCESSSTALEMATE( $n, n_t$ )
13:            $n.cross \leftarrow 0$   $\triangleright$  Reset counter
14:         end if
15:       end if
16:     end if
17:   end for
18:   return  $\epsilon_t$ 
19: end procedure

```

---

#### 4.4.4 Translating Rivers

For all nodes  $n_{n_x r}$  that cross a river,  $r$ , we compute an average vector  $v_{avg}$  used to translate  $r$ . Whenever a node,  $n$ , crosses a river, we store a vector  $\overrightarrow{nn_t}$  that points in the direction of the translation. We then use a heuristic to translate  $r$  using the

---

**Algorithm 3** Procedure to test if a node's translation path,  $\overline{\mathbf{nn}_t}$  intersects a river.

---

**Input:**

$\overline{\mathbf{nn}_t} \leftarrow$  the node's translation path

$\mathbf{r} \leftarrow$  a river feature

**Output:**

Returns *True* if the node crosses a river.

**Local variables:**

$b_{\overline{n}}, b_{\overline{e}} \leftarrow$  the bounding boxes for  $\overline{\mathbf{nn}_t}$  and  $\overline{e}$

$\overline{e} \leftarrow$  an edge of  $\mathbf{r}$

```

1: procedure TESTINTERSECTION( $\overline{\mathbf{nn}_t}, \mathbf{r}$ )
2:   for each  $\overline{e} \in \mathbf{r}$  do
3:      $b_{\overline{n}} \leftarrow$  GetBoundingBox ( $\overline{\mathbf{nn}_t}$ )
4:      $b_{\overline{e}} \leftarrow$  GetBoundingBox ( $\overline{e}$ )
5:     if  $b_{\overline{n}}$  intersect  $b_{\overline{e}} = \text{True}$  then
6:       return  $\overline{\mathbf{nn}_t}$  intersect  $\overline{e}$ 
7:     end if
8:   end for
9:   return False
10: end procedure

```

---

average vector of node intersection

$$\mathbf{v}_{avg} = \sum_{i=1}^{\epsilon_t} \frac{\overrightarrow{\mathbf{n}_i \mathbf{n}_{i_t}}}{\epsilon_t}$$

, where  $\epsilon_t$  is the number of nodes intersecting the river. This step intends to create space for the next iteration of node translation without crossing a river. The detailed procedure for translating rivers is provided by [Algorithm 4](#).

When a river,  $\mathbf{r}$ , is translated by  $\mathbf{v}_{avg}$ , this can trigger a scenario where nodes are crossed by a translated river, denoted  $\mathbf{n}_{r_x n}$ . As a heuristic, we also translate these nodes by  $\mathbf{v}_{avg}$ . The reasoning behind this is that  $\mathbf{v}_{avg}$  indicates which direction the river needs to be translated to create space for the nodes that are too crowded together. In practice,  $\mathbf{v}_{avg}$  is multiplied by a scaling factor  $\propto \mathbf{v}_{avg}$ . Thus, we can influence how far  $\mathbf{r}$  is translated in each iteration of the layout algorithm. We can use this to ensure smooth transitions between iterations.

---

**Algorithm 4** Procedure to translate rivers.

---

**Input:** $L \leftarrow$  a list of  $\mathbf{n}$  representing regions**Local variables:** $R \leftarrow$  a list of  $\mathbf{r}$  representing river features $\mathbf{n} \leftarrow$  a node is an object with the following properties: $\mathbf{n}.x, \mathbf{n}.y$ , or  $\mathbf{n}(x, y) \leftarrow$  the x and y coordinates of  $\mathbf{n}$  $\mathbf{n}_p \leftarrow$  the previous position of  $\mathbf{n}$  $\mathbf{n}_t \leftarrow$  the translated position of  $\mathbf{n}$  $\epsilon_t \leftarrow$  the number of nodes intersecting  $\mathbf{r}$ 

```
1: procedure TRANSLATERIVER( $L$ )
2:   for each  $\mathbf{r} \in R$  do
3:      $\epsilon_t \leftarrow 0$ 
4:      $\vec{v}_{\mathbf{r}} \leftarrow (0, 0)$   $\triangleright$  Hold the sum of vectors  $\overrightarrow{\mathbf{n}\mathbf{n}_t}$ 
5:     for each  $\mathbf{n} \in L$  do
6:       if  $\mathbf{n}(x, y) \neq \mathbf{n}_t(x, y)$  then
7:         if TESTINTERSECTION( $\overrightarrow{\mathbf{n}\mathbf{n}_t}, \mathbf{r}$ ) = True then
8:            $\vec{v}_{\mathbf{r}} \leftarrow \vec{v}_{\mathbf{r}} + \overrightarrow{\mathbf{n}\mathbf{n}_t}$ 
9:            $\epsilon_t \leftarrow \epsilon_t + 1$ 
10:        end if
11:      end if
12:    end for
13:    Translate river  $\mathbf{r}$  by the average vector  $\frac{\vec{v}_{\mathbf{r}}}{\epsilon_t}$ 
14:  end for
15: end procedure
```

---

#### 4.4.5 Process Stalemates

As the FNOR always attempts to produce an optimal node layout where node distribution and translation are minimised, a node's translation path can repeatedly intersect a river due to congestion, creating a stalemate situation, as shown in [Figure 4.5](#). If a node is translated between two positions,  $\mathbf{n}$  and  $\mathbf{n}_t$ , for  $w$  iterations (a user-adjustable parameter), we introduce a heuristic solution: constructing a corridor to alleviate congestion. A corridor,  $c$ , is a rectangle with a width of  $c_w$  and a length of  $c_l$ , formed by deriving two edges  $\overline{e_{p^1}}$  and  $\overline{e_{p^2}}$  such that  $\overline{e_{p^1}} \parallel \overline{e_{p^2}} \parallel \overline{\mathbf{n}_t\mathbf{n}_c}$  (See [Figure 4.6C](#) and [D](#)). All nodes enclosed by  $c$  are then translated by  $\overrightarrow{\mathbf{n}_t\mathbf{n}}$  to alleviate the congestion (See [Figure 4.6E](#)). The procedure for constructing corridors is provided by [Algorithm 5](#).

---

**Algorithm 5** Procedure to derive a corridor to resolve stalemates. We use an SVG canvas, where the point of origin (0,0) is located at the top left corner, with the x-axis extending to the right and the y-axis extending downwards (See Figure 4.5).

---

**Input:**

$\mathbf{n} \leftarrow$  the node used to derive the corridor

**Global variables:**

$c_l \leftarrow$  the length of a corridor

$c_w \leftarrow$  the width of a corridor

**Local variables:**

$c \leftarrow$  the corridor

$\mathbf{n}_c \leftarrow$  the point extending  $\overrightarrow{\mathbf{n}_t \mathbf{n}}$  such that  $|\mathbf{n}_t \mathbf{n}_c| = c_l$

$\overline{e_{p^1}}, \overline{e_{p^2}} \leftarrow$  the edges parallel to  $\overrightarrow{\mathbf{n}_t \mathbf{n}_c}$

$corridor \leftarrow$  a rectangle formed by  $\overline{e_{p^1}}$  and  $\overline{e_{p^2}}$

```

1: procedure PROCESSSTALEMATE( $\mathbf{n}$ )
2:    $\mathbf{n}(x, y) \leftarrow \mathbf{n}_p(x, y)$ 
3:    $\mathbf{n}_c \leftarrow \text{DERIVEPOINT}(\overrightarrow{\mathbf{n}_t \mathbf{n}}, c_l)$ 
4:    $\overline{e_{p^1}} \leftarrow \text{DERIVEPARALLELEDGE}(\overrightarrow{\mathbf{n}_t \mathbf{n}_c}, \frac{c_w}{2})$ 
5:    $\overline{e_{p^2}} \leftarrow \text{DERIVEPARALLELEDGE}(\overrightarrow{\mathbf{n}_t \mathbf{n}_c}, -\frac{c_w}{2})$ 
6:    $c \leftarrow \begin{bmatrix} \overline{e_{p^1}}.start & \overline{e_{p^1}}.end \\ \overline{e_{p^2}}.start & \overline{e_{p^2}}.end \end{bmatrix}$ 
7:   for each  $\mathbf{n}_{in}$  inside  $c$  do
8:      $\overrightarrow{\mathbf{n}_{in} \mathbf{n}_{in_t}} = \overrightarrow{\mathbf{n}_t \mathbf{n}}$ 
9:      $\mathbf{n}_{in}(x, y) \leftarrow \mathbf{n}_{in_t}(x, y)$ 
10:  end for
11: end procedure

```

---

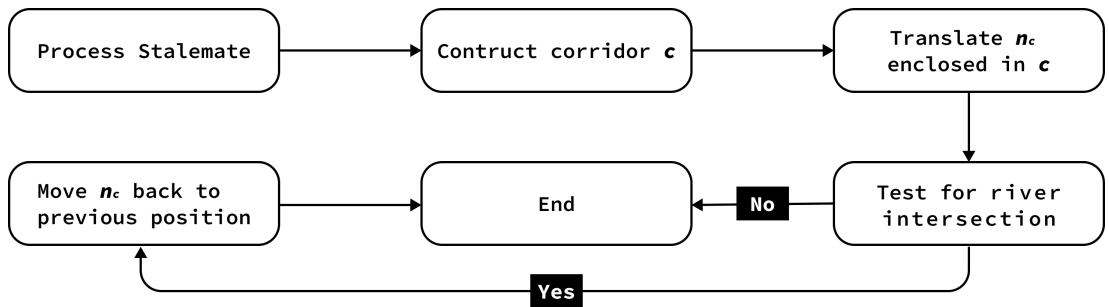


Figure 4.4: A flowchart illustration of stalemate processing. See Section 4.4.5, Figure 4.5, and refer to our video demonstration (from 1:27 to 1:46) for details.

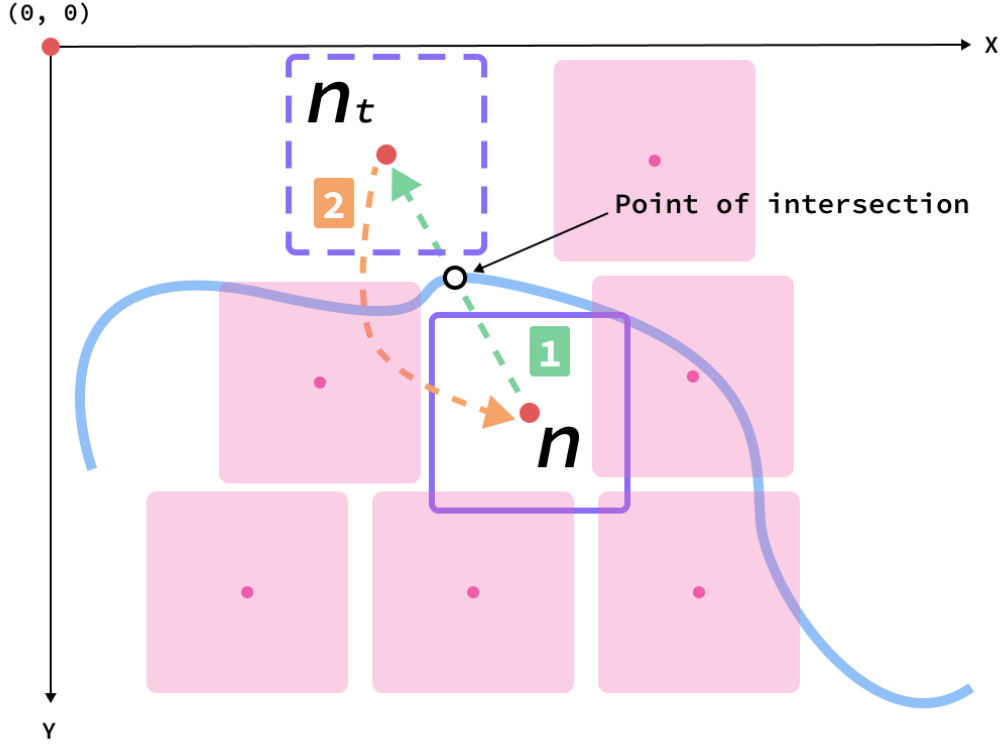


Figure 4.5: A stalemate: when a node's translation path  $\overrightarrow{nn_t}$  (iteration 1) intersects a river  $w$  times. The node is translated back to its previous position (iteration 2). A stalemate can occur when the area is congested and the node cannot translate to a new position without intersecting a river.

#### 4.4.6 Terminating the Algorithm

The algorithm terminates when  $\epsilon_c$  reaches  $\epsilon_{c_{max}}$ , the error tolerance set by the user.

We adopt the maximum cartographic error from Alam et al., namely:

$$\epsilon_{c_{max}} = \max_{n \in L} \frac{|n_i - n_{i_t}|}{\max\{n_i, n_{i_t}\}}$$

where  $n_i$  and  $n_{i_t}$  are the initial and translated regions in the cartogram,  $L$  represent the list of regions, and  $\epsilon_{c_{max}}$  is a normalised value that we express as a percentage. For a detailed derivation of the formula, refer to the work by Alam et al. Because the algorithm processes node-river intersections, we can measure a novel kind of error, namely, topological error  $\epsilon_t$ . We maintain  $\epsilon_t = 0$ , however, we can count how many nodes would have crossed a river if we did not test for this and simply let nodes cross rivers. We express  $\epsilon_t$  in the normalised range  $\epsilon_t \in [0, 1]$ , where  $\epsilon_{t_{max}}$  is the case where all nodes cross a river.



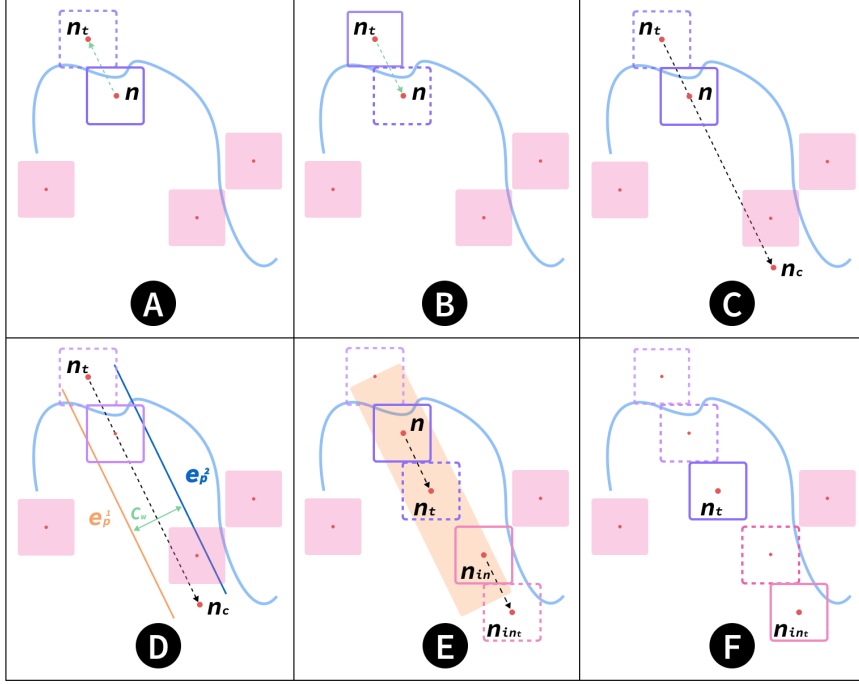


Figure 4.6: A stalemate occurs when a node’s translation path  $\overrightarrow{nn_t}$  intersects a river for  $w$  times, as shown in (A). To address this, we derive a corridor (orange rectangle in (E)) based on  $n$  and  $n_t$ . All nodes within the corridor are translated based on  $\overrightarrow{n_t n}$ , such that  $\overrightarrow{nn_t} = \overrightarrow{n_{in} n_{in_t}}$ . For clarity in the illustration, we place nodes sparsely in this figure.

When the algorithm terminates, the node layout is considered optimal where no nodes have crossed or crossed a river (denoted  $\epsilon_t = 0$  and  $\epsilon_c < \epsilon_{c_{max}}$ ). Every node remains on the same side of the river as its centroid.

#### 4.4.7 User Options

Figure 4.7 presents an overview of the application, including user options. The user can adjust the following parameters:

##### Rendering Visibility:

The rendering visibility of various elements, including the choropleth shapefile, rivers, nodes, and node centroids, can be toggled on and off.

##### Node Mapping:

Both size and colour of the nodes can be mapped to different EHR attributes or

set to uniform.

#### **Overlap Removal Speed:**

The overlap removal process can be observed step by step, or the algorithm can be run automatically.

#### **Maximum Cartographic Error:**

The user can adjust the maximum cartographic error,  $\epsilon_{c_{max}}$ , which is used to terminate the algorithm.

#### **River Translation:**

The behaviour of rivers during the process can also be adjusted by the user: 1) Enable river crossing: this option allows nodes to cross rivers. Nodes cannot cross rivers by default; 2) Disable river translation: this option disables the translation of rivers, rivers are translated by default. Both options are useful for generating different layouts and to observe the behaviour of the hybrid layout algorithm.

#### **Corridor Length:**

The user can define the length of a corridor that is used to resolve stalemate situations. A longer corridor length allows more nodes to translate during a stalemate. The default corridor length is three times the max node size.

#### **River Thickness and Resolution:**

Increasing the thickness of rivers may improve the recognisability of the cartogram. Similarly, increasing the resolution of rivers, at the expense of the speed of node-river intersection test, may produce a layout with higher legibility.

## **4.5 User Evaluation**

We conduct a user study to evaluate the effectiveness of our approach. The evaluation is designed to test the hypothesis that the introduction of rivers can improve the legibility and recognisability of cartograms. We chose this type of evaluation because legibility is a human-centred characteristic. Many different types of statistical error metrics have been evaluated in previous work [176], however, our focus is more user-centric in nature.

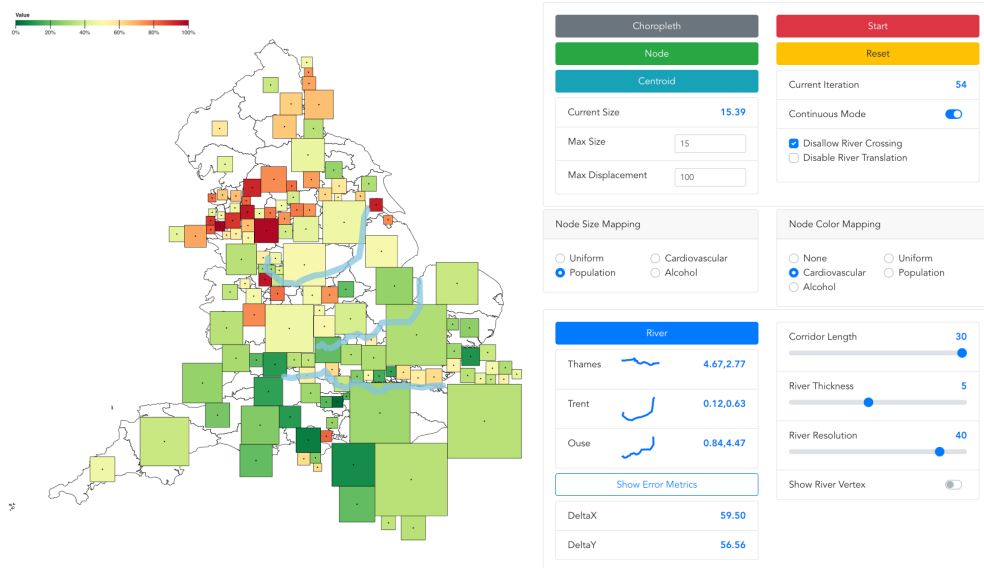


Figure 4.7: A screenshot of the user interface. User options are provided to adjust the terminating conditions (size and error), colour mapping, and visibility of nodes and rivers. Other options include the ability to control the overlap removal behaviour: rivers can be static or dynamic. See [Section 4.4.7](#) for more on user options. In this figure,  $\epsilon_{c_{max}} = 1.875\%$ , and an  $\epsilon_t$  of 9.631% is eliminated.

### 4.5.1 Study Hypothesis

Our work builds on the results of the previous study [218], which demonstrated that the incorporation of topological characteristics in cartograms improves recognisability. Unlike the earlier approach, which uses static topological features, our method introduces dynamic topological features. Thus, we hypothesise that the introduction of rivers improves cartogram legibility and recognisability, specifically:

**H1:** The presence of rivers in cartograms will improve participants’ ability to correctly locate the target CCG as measured by accuracy.

**H2:** The presence of rivers in cartograms will decrease response times when identifying the target CCG.

We tested these hypotheses using location tasks, in which participants matched a target CCG from a standard choropleth (left) to its counterpart in a cartogram generated using our approach (right). See [Figure 4.8](#).

### 4.5.2 User Study Variables

In this section, we discuss the variables of our user study.

#### Independent Variables

*River presence:* whether rivers are displayed, their presence impacts the interpretation of the cartogram.

*Node river crossing:* whether nodes are allowed to cross rivers, this affects the final layout of the cartogram.

### **Dependent Variables**

*Accuracy:* Given a CCG location in the choropleth on the left, we asked the participant to locate the corresponding node in the cartogram. Accuracy as the primary dependent variable is measured by the percentage of correctly identified target CCGs out of total attempts.

*Response time:* Another dependent variable is the time it takes the participants to complete each task.

### **Control Variables**

*Choice of Colormap:* We use D3’s built-in interpolateRdYlGn (red-yellow-green) colormap to depict the data in our cartograms.

*Target CCG Communication:* We inform the participant about the target CCG they need to find. The target CCG blinks (between its original colour and black) every 2 seconds to ensure visibility.

## **4.5.3 User Study Design**

In this section we describe the user study participants, data sets, and the experimental procedure. To test our hypothesis that the inclusion of rivers improves the legibility and recognisability of Demers cartograms, we designed a user study to evaluate the impact of river presence and node river crossing on participants’ performance in identifying CCGs. This study is particularly relevant to EHR visualization, as accurate and efficient interpretation of geospatial health data (e.g., CCG-level health outcomes) is critical for clinicians and policymakers to make informed decisions. We employed a within-subjects design, where each participant completed all tasks under all four combinations of the two independent variables. The study was approved by the University of Nottingham’s Research Ethics Committee (Ref: 2021-2022-001).

## Participants

We recruited 24 participants.

- Gender: 10 females and 14 males
- Age Group: 18-24 (16), 25-29 (8)
- Education: 1 Ph.D., 11 Master's, 8 Bachelor's, 4 Others

Participants with colour deficiency were not recruited to ensure accurate interpretation of colour-coded elements in the visualisation. This criterion was established to maintain consistency in evaluating the effectiveness of the proposed visualisations and avoid potential biases caused by difficulties in distinguishing colours.

## Data sets

We used the following EHR data sets from NHS Digital [\[351\]](#) for our evaluation:

- Population
- Under 75 mortality from cardiovascular disease
- Emergency admissions for alcohol-related liver disease
- Alcohol-specific admission and readmission

See [Section 4.3.3](#) for a detailed description of the data sets.

A choropleth map displaying 135 CCGs was rendered on the left side of the screen, while a cartogram-based view was rendered on the right. In both views, the colour was assigned to the percentage of the condition in the data set. See [Figure 4.8](#) for an example.

## Procedure

Due to pandemic restrictions, the user study was designed to be conducted online. Participants used their own retail hardware, devices with small screens were not recommended due to the large size of the rendered cartograms. We provided detailed instructions to use a screen resolution of 1920x1080 pixels and to complete the study in a distraction-free environment to mitigate variability. However, differences in device hardware and internet connectivity may have introduced minor variability, which we discuss in [Section 4.6](#).



Figure 4.8: A sample location task for participants. The left shows the choropleth map, and the right shows the corresponding cartogram. Both images show the three longest rivers in England, with the size and colour of nodes representing the prevalence of the selected disease. The target CCG blinks on the choropleth (shown in black), and participants are asked to identify this CCG on the cartogram. In this figure,  $\epsilon_{c_{max}} = 1.875\%$ , and an  $\epsilon_t$  of 6.667% is eliminated.

The user study included four parts:

**P1 Training Session:** The participants were instructed to complete a training session, available in both text and video formats (available at <https://tinyurl.com/demerscartogram>). The session introduced key concepts such as choropleths and cartograms used in the tasks. It also provided a demonstration of the tasks, guidance on navigating the user interface, and a brief overview of the data sets used in the study.

**P2 Practice Session:** The participants were then given three practice tasks to familiarise themselves with the user study design. A sample task was shown in Figure 4.8, and accessible at <https://ghr.wangqiru.com/#/P1> and included in the online materials available at <https://osf.io/q39w7> and <https://github.com/thevisgroup/Demers-Cartogram-with-Rivers>. The practice tasks were identical to the actual tasks, accompanied by instructions provided before the actual tasks commence. These tasks ensured that participants understood how to interact with the visualisation and have a similar level of familiarity with the interface. The results from these practice tasks were not included in the final analysis. In the instructional video, a demonstration of three sample tasks was also included.

**P3 Main Session:** The participants were instructed to complete 16 location

tasks that involve 4 target CCGs: *E38000243, NHS Nottingham and Nottinghamshire, E38000136, NHS Oxfordshire, E38000242, NHS Northamptonshire, and E38000244, NHS South East London*. The 4 CCGs were carefully selected to avoid extreme cases such as very large or very small areas, highly distinct or barely distinguishable colours, and locations near map edges, to avoid potential biases. All participants completed the same set of tasks, each presented under four different conditions, to enable within-subject comparison while avoiding repetition of task content. Each task required the participant to identify a target CCG on the choropleth map (left) and locate its corresponding representation in the cartogram (right). The 16 tasks were presented in a fixed block order, grouped by condition shown in [Table 4.5](#). Each block corresponded to one combination of river presence and node river crossing ability, with 4 tasks per block using different CCG locations. This fixed order was chosen to standardise the test sequence across participants while varying the content to mitigate task-specific learning effects.

Accuracy and response time were recorded.

**P4 Post-Study Session:** The participants were asked to complete a questionnaire that consists of 5-Point Likert scale questionnaire, as shown in [Section 4.8.2](#).

#### 4.5.4 User Study Analysis

This section reports the results of the user study evaluating the effects of river presence and node river crossing ability on task accuracy and response time. We conducted a two-way repeated-measures experiment with two within-subjects factors: river presence (present vs. absent) and node river crossing (allowed vs. disallowed). Participants completed location tasks under different conditions and their accuracy and response times were analysed. The results indicate that river presence significantly improves accuracy, while response times are influenced by an interaction between river presence and node river crossing ability.

#### Data Cleaning and Preprocessing

Our study recruited 24 participants, each assigned to complete 16 location tasks, resulting in a total of 384 responses. To improve data reliability, responses with extreme

	Original Results	Outliers Removed
Count	384	346
Mean	13,830.05	8,578.26
Median	7,944	7,521
Range	1,783 - 177,423	1,783 - 24,420
IQR	7,918	6,086

Table 4.4: The table compares various statistical measures of response times in milliseconds (ms), showing the original values versus those after removing outliers. Outliers were defined as values above  $Q3 (12,839 \text{ ms}) + 1.5 \times \text{IQR} (24,716 \text{ ms})$ .

	Condition 1	Condition 2	Condition 3	Condition 4
River Presence	T	T	F	F
Node River Crossing	F	T	F	T
Total Responses	91	84	83	88
Correct Responses	58	50	40	36
Accuracy (%)	63.74	59.52	48.19	40.91
Mean Response Time (ms)	7686.38	9176.75	8657.90	8854.15

Table 4.5: The table presents the accuracy and mean response times (in milliseconds) of four different conditions, after removing outliers.

response times were removed using an upper-bound threshold based on the interquartile range (IQR). Response times greater than the upper quartile plus 1.5 times the IQR were excluded as outliers ( $n = 38$ ), resulting in 346 valid responses. These responses were assumed to reflect lapses in participant attention or technical delays. See [Table 4.4](#) for a comparison of statistical measures before and after outlier removal.

## Accuracy Analysis

[Table 4.5](#) shows the accuracy measured by the number of correct CCGs chosen by the participants. Conditions without rivers (Conditions 3 and 4) exhibited lower accuracy (48.19% and 40.91%, respectively) compared to those with rivers (63.74% and 59.52%).

Accuracy data were analysed using a binomial generalised linear mixed model (GLMM), with river presence and node river crossing ability as fixed effects and a random intercept for each participant.

The model revealed a significant main effect of river presence ( $OR = 2.124$ ,  $p = 0.015$ ), indicating that the odds of a correct response were significantly higher when rivers were present. However, this effect should be interpreted with caution due to the fixed block order and repeated use of CCG targets, which may have introduced order-



related confounds. No significant effects were found for node river crossing ability ( $OR = 1.344$ ,  $p = 0.339$ ) or for the interaction between the two factors ( $OR = 0.889$ ,  $p = 0.789$ ). The random intercept term for participants showed negligible variance, indicating no between-subject variability in accuracy, likely due to the standardised task design. Nevertheless, we retained the random intercept to respect the repeated-measures structure of the study design and to avoid anti-conservative inference.

Overdispersion was assessed by computing the Pearson chi-squared statistic divided by the residual degrees of freedom. The dispersion parameter ( $\chi^2 = 346.000$ ,  $df = 342$ ,  $\hat{\Phi} = 1.012$ ) indicates no evidence of overdispersion and consistency with binomial model assumptions.

## Response Time Analysis

Response time data were analysed using a linear mixed-effects model (LMM) to account for within-subject variance, with a random intercept for each participant.

There was no significant main effect of river presence ( $\beta = 374.841$ ,  $p = 0.554$ ) or node river crossing ability ( $\beta = -180.561$ ,  $p = 0.776$ ). The interaction term between river presence and node river crossing ability showed a trend toward significance ( $\beta = -1,449.290$ ,  $p = 0.105$ ), suggesting a potential interaction pattern where node river crossing ability reduces response time when rivers are present, a pattern consistent with previous findings, although the interaction did not reach statistical significance in the current sample. As block order and task content were not counterbalanced, this observed trend may also be partially influenced by order or fatigue effects.

Although none of the fixed effects reached conventional significance levels, we also fitted a model using log-transformed response times to account for the right-skewed distribution. The transformed model revealed no significant interaction between river presence and node river crossing ability ( $\beta = -0.138$ ,  $p = 0.173$ ), consistent with the original analysis. Main effects were not statistically significant. Residual diagnostics were then conducted on the transformed model. The histogram of residuals showed an approximately symmetric, bell-shaped distribution, and the Q-Q plot indicated that the residuals closely followed the theoretical normal distribution. This indicates that the primary finding is robust to distributional assumptions.

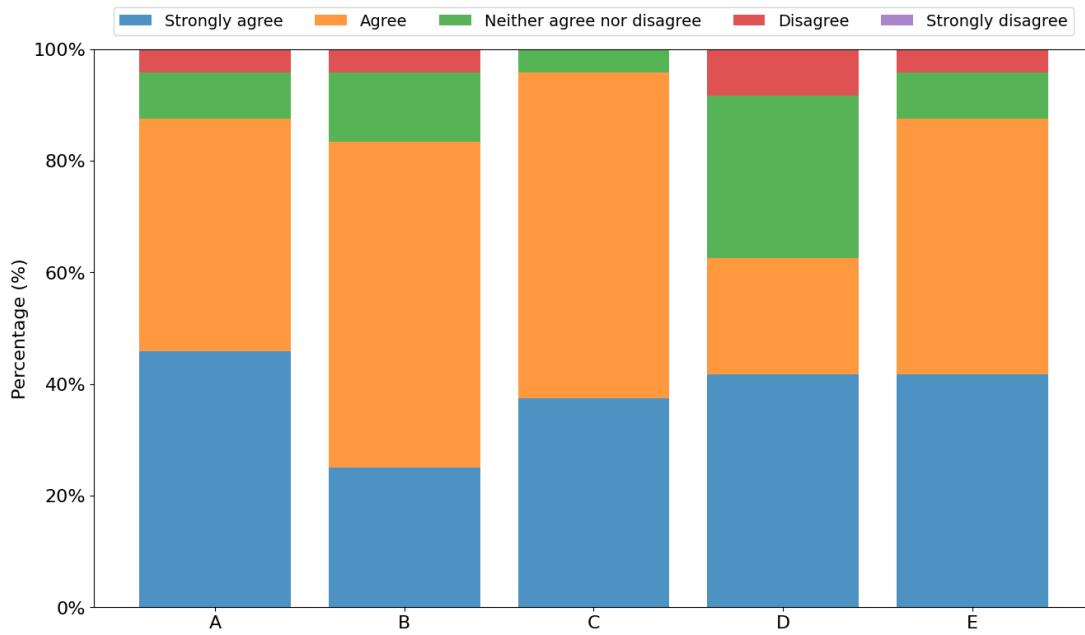


Figure 4.9: The stacked bar chart shows the user study participant responses of Likert Scale questions.

### Likert Scale Analysis

Figure 4.9 shows the results of the following Likert Scale questions:

- (A) 87.50% of participants agree that including rivers is useful.
- (B) 83.33% of participants agree that rivers increase the legibility of a cartogram.
- (C) 95.83% of participants agree that including rivers makes cartograms easier to understand.
- (D) 62.50% of participants agree that including rivers makes CCGs easier to locate.
- (E) 87.50% of participants agree that including rivers adds value to the standard cartogram.

The responses indicate a strong subjective preference for cartograms with rivers, aligning with the objective accuracy improvements observed in the study. The fact that nearly all participants agreed that rivers made cartograms easier to understand (95.83%) and improved overall readability (83.33%) is consistent with the accuracy findings, where the presence of rivers significantly increased the number of correctly identified CCGs.

However, a discrepancy arises when comparing Likert scale responses with response time results. While most participants (87.50%) agreed that rivers added value to the cartogram, the response time analysis suggests that their benefit was conditional on

node river crossing ability, and rivers alone do not universally speed up decision-making, instead, their impact depends on how the layout allows users to process spatial information. The 62.50% agreement that rivers helped locate specific CCGs suggests that while they improve overall comprehension, their influence on precise localisation is somewhat weaker. This aligns with the finding that accuracy improvements were clear, but response times were not universally reduced.

Together, the statistical and subjective results demonstrate that rivers enhance cartogram legibility, particularly in flexible layouts, while accuracy benefits appear more robust across conditions.

In summary, H1 is partially supported by the data, the presence of rivers moderately improved participants' accuracy in identifying the correct CCG. H2 is partially supported by the data, while no significant main effect of river presence on response time was found, a significant interaction between river presence and node river crossing ability indicated that the benefit of rivers for efficiency was conditional on node river crossing ability. This suggests that rivers can enhance decision-making speed, but only in less constrained spatial configurations.

## Discussion

The result of our study provides support for our primary hypothesis that rivers enhance cartogram legibility and recognisability. The presence of rivers significantly increased participants' accuracy in correctly identifying the target CCGs, with participants being more than twice as likely to respond correctly when rivers were present ( $OR = 2.124$ ,  $p = 0.015$ ). Conditions where rivers were present resulted in higher accuracy rates (63.74% and 59.52%) compared to conditions without rivers (48.19% and 40.91%).

The presence of rivers improved task accuracy, suggesting that the inclusion of familiar geographical features may have supported users in identifying correct CCGs. By contrast, neither node's ability to cross rivers nor its interaction with river presence significantly affected accuracy. This outcome implies that restricting layout flexibility did not affect the participants' ability to identify target regions.

Response time analysis revealed more nuanced effects. Response times were the lowest when rivers were present and nodes were not allowed to cross rivers (condition 1:

7,686.38 ms). This pattern was supported by the interaction effect in the mixed-effects model, which showed a trend towards faster responses when rivers were present and nodes cannot cross rivers ( $\beta = -1449.290$ ,  $p = 0.105$ ), though the effect did not reach statistical significance. In contrast, when nodes are allowed to cross rivers, the presence of rivers slightly increased response times, potentially due to increased cognitive load associated with interpreting a more visually complex layout.

Taken together, these results partially support the second hypothesis. The presence of rivers did not universally accelerate task completion, but their contribution to efficiency was conditional on node river crossing ability.

Finally, the subjective responses captured in the Likert-scale questions reinforce the quantitative results. Participants overwhelmingly agreed that rivers improved cartogram legibility and readability. However, only 62.50% felt that rivers made it easier to locate specific CCGs, which may reflect the subtle or conditional nature of their influence on response time.

## 4.6 Limitations and Future Work

This section discusses the limitations of our approach and potential directions for future research. Given the novelty of integrating topological features into Demers cartograms, there are several areas for improvement and extension.

### 4.6.1 Experimental Design Confounds

A limitation of the study design is that the four task blocks (conditions) were presented in the same fixed order for all participants, and each block reused the same four target CCGs. This introduces potential confounds due to order effects (e.g., fatigue or learning) and repeated content exposure. For example, the significant accuracy advantage observed in river-present conditions may partially reflect their earlier presentation in the task sequence. Without counterbalancing, it is not possible to disentangle these effects. Future studies should randomise both block order and CCG assignment to control for such confounds.

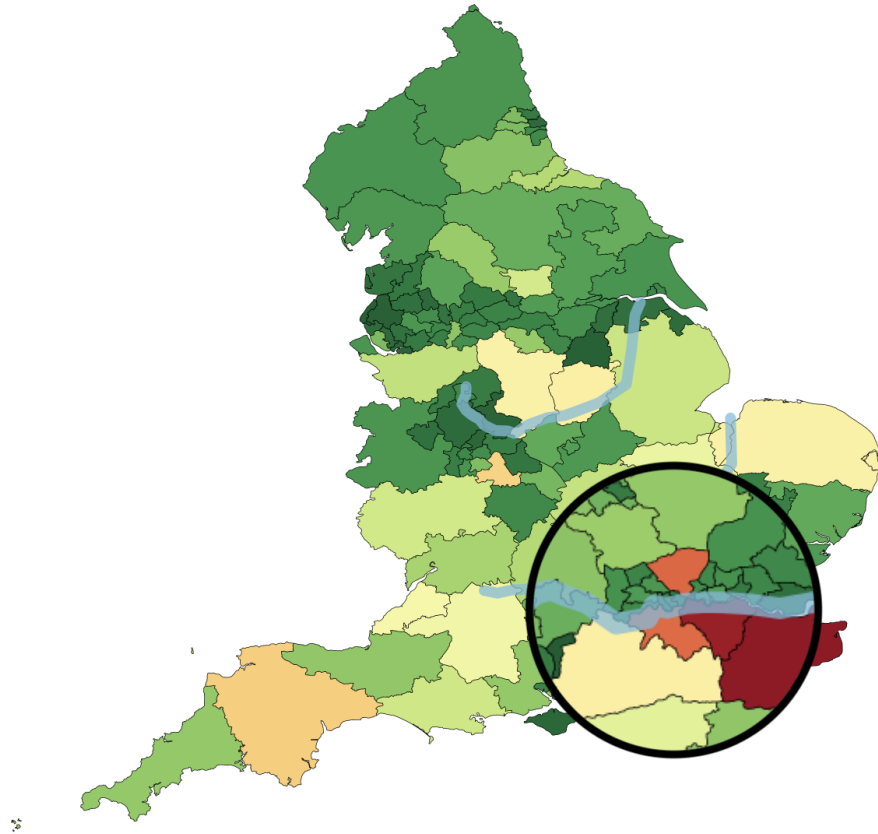


Figure 4.10: Due to colour and relative location, we believe the CCGs in the black circle are easier to locate.

### 4.6.2 Colormap choice

The first limitation is the colormap that we use to depict the data in our cartograms. We use D3’s built-in `interpolateRdYlGn` colormap, a diverging colour scheme of red, yellow, and green. While effective in many cases, alternative colormaps may offer better differentiation between high and low values. Future research should investigate different colormap options and their impact on interpretability, including testing perceptually uniform colormaps and those optimised for colour deficiencies. In the user study, we carefully avoid extreme values where the location or colour of the CCG makes it easier to locate the target. See [Figure 4.10](#) for an example. We plan to explore the impact of different colormaps on the legibility of cartograms in future work.

### 4.6.3 Overlap removal algorithm choice

Another limitation is the algorithm (FNOR) we use. A future enhancement would involve designing an overlap removal algorithm that inherently incorporates topological

constraints, potentially reducing computational complexity and ensuring more efficient layouts. Currently, the runtime of our layout algorithm is approximately 30 milliseconds for each iteration. When the quantity of nodes and features increases, generating the optimal layout demands several hundred iterations.

#### **4.6.4 Generalisability**

Future work also includes generalisations and extensions of the algorithm, e.g., the use of other features in the cartogram layout such as additional rivers, major highways, lakes, and coastlines, etc. We also consider whether increasing the length of the rivers as the size of the nodes increases would be a useful option. We would like to explore the case of river-river intersections (or confluence) and test the approach on more diverse geographic regions, such as North America and Europe, to assess its generalisability beyond the UK. We also considered the idea of deforming the rivers as part of the layout algorithm, however, this idea is open to future work.

#### **4.6.5 Improved User Study**

The user study was conducted in an online environment due to pandemic restrictions, leading to variations in screen sizes, hardware specifications, and internet connections. Future studies should be conducted in controlled settings to ensure consistency in display quality and user interaction. Future studies should also randomize task order or use a counterbalanced design. It is important to acknowledge that our layout algorithm itself was not explicitly evaluated as part of this study. Future studies could introduce direct evaluation metrics for the layout algorithm. For example, these could include user comparisons of different algorithm-generated layouts for the same data set.

### **4.7 Conclusions**

The work presented in this chapter advances the field of spatial visualisation in the context of EHR Vis by refining the Demers Cartogram with a novel river-preserving technique. We first propose a new algorithm to generate cartograms with rivers, and then present a prototype to support the exploration of cartograms with rivers. The pro-

posed approach enhances the interpretability of spatial health trends while preserving critical geographical features, ensuring that users, including healthcare professionals and policymakers, can make informed decisions based on geographically coherent representations of patient data. Through a user-centred evaluation, we demonstrated that the presence of rivers enhances cartogram legibility and recognisability. The results indicate that users were more accurate in identifying regions when rivers were present, but the lack of counterbalancing in task block order means this finding should be interpreted with caution. Response time improvements also depended on river crossing ability of the node, suggesting a trade-off between cognitive load and spatial clarity. Subjective responses from participants further corroborated these findings, with strong agreement that rivers enhanced the usefulness, legibility, and clarity of cartograms. Future work should explore additional constraints, optimise computational performance, and conduct further validation studies in controlled environments with more participants.

This work aligns with and contributes to the broader “EHR Vis” framework by emphasising the importance of geospatial context in EHR analysis. While [Chapter 3](#) focuses on structuring and abstracting unstructured text-based EHR data into meaningful representations, this chapter demonstrates how structured geospatial data can be made more comprehensible through visual representations to support EHR analysis and decision-making.

## 4.8 Appendix

### 4.8.1 Preprocessing Shapefiles

Shapefiles from different sources are likely to be incompatible. In our case, the NHS CCG shapefile is incompatible with the river shapefiles. The major reason for the incompatibility is the coordinate reference system (CRS). The CRS of the CCG shapefile is EPSG:27700 (OSGB36 - British National Grid). The CRS of the river shapefiles is EPSG:4326 (WGS84 - World Geodetic System). Here, we provide some preprocessing steps using QGIS (version: 3.26.0-Buenos Aires) [[355](#)] to handle the incompatibility and reduce the shapefile size to improve performance.

## Import Shapefiles into QGIS

We first load all three river shapefiles into QGIS [Figure 4.11](#), followed by the CCG shapefile [Figure 4.12](#).

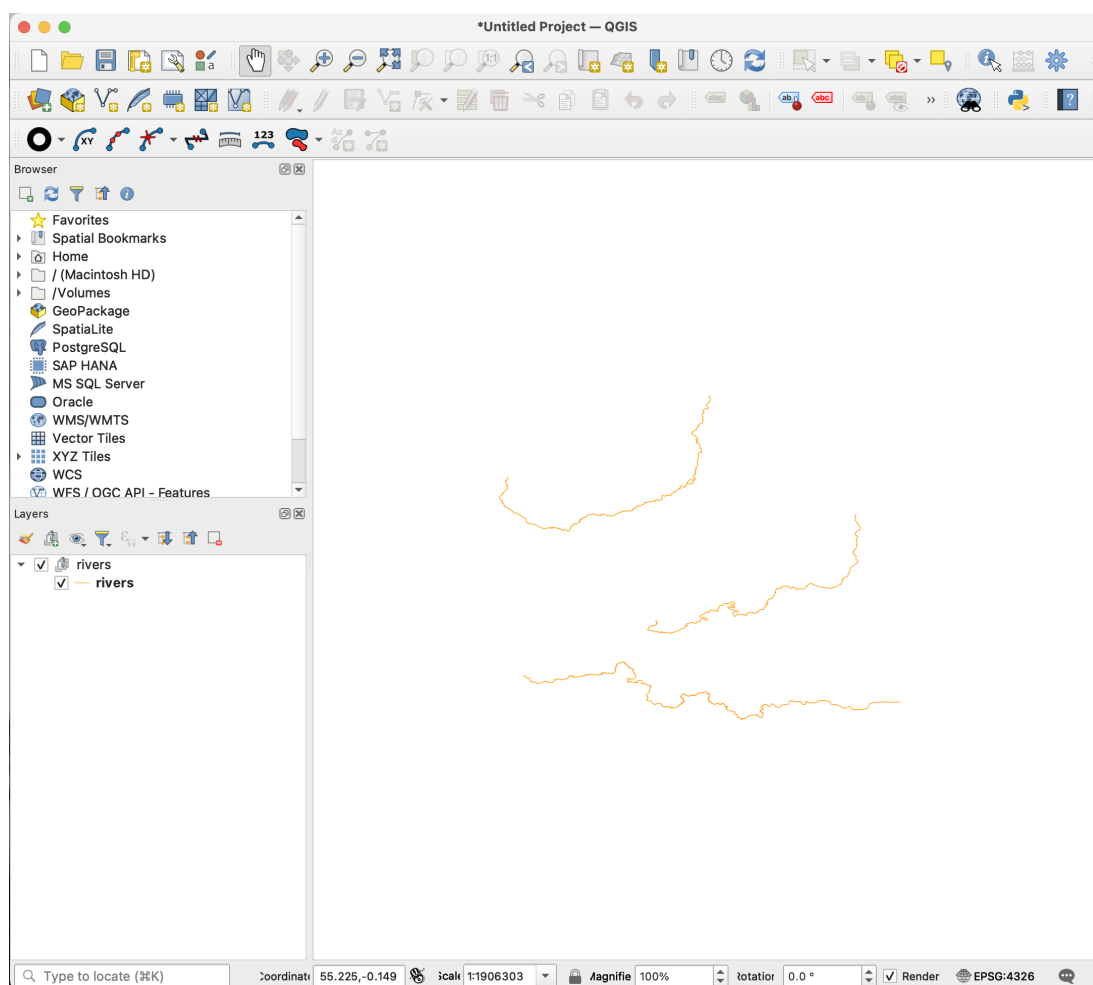


Figure 4.11: QGIS interface, with River Trent, River Great Ouse, and River Thames (from top to bottom) imported.



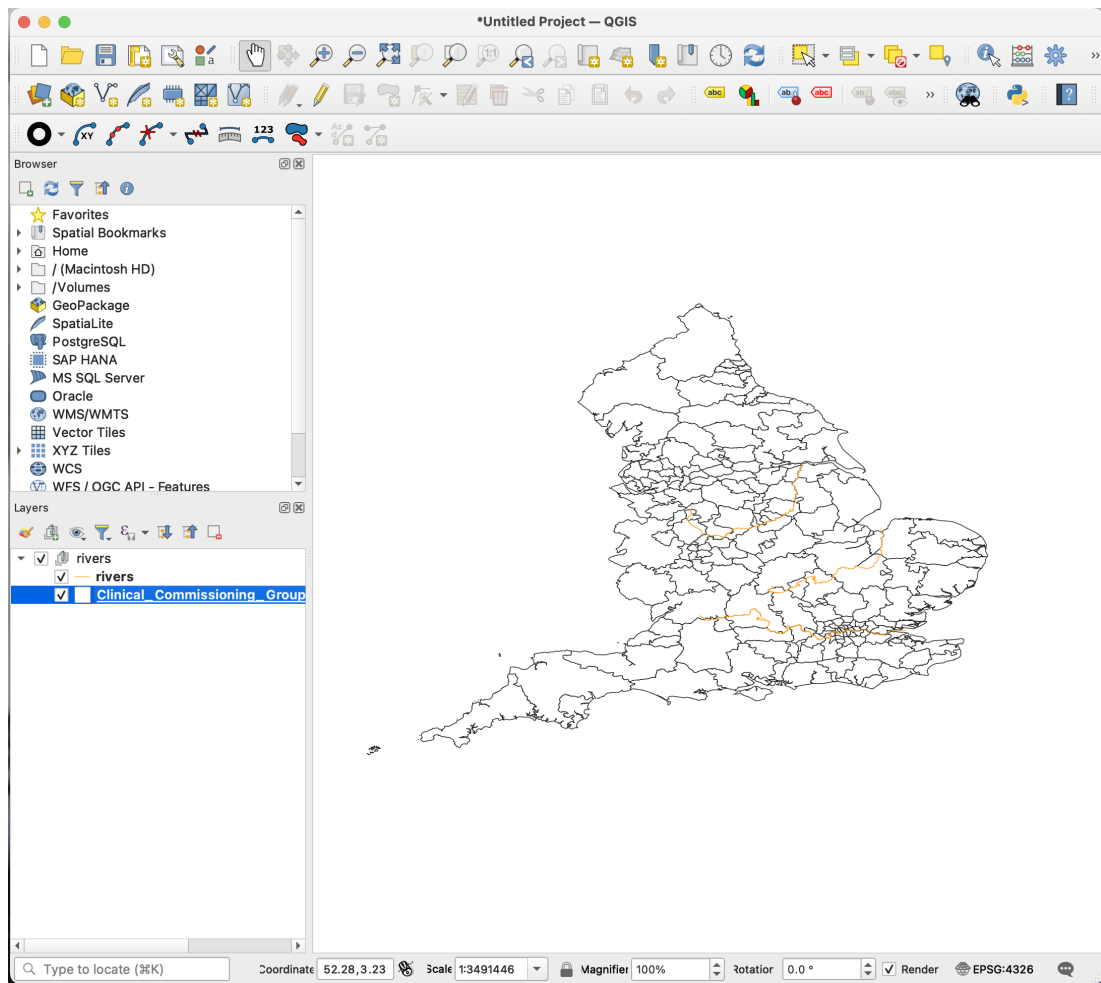


Figure 4.12: QGIS interface, with all NHS CCGs imported.

## Export Shapefiles in GeoJSON and Unify the Coordinate Reference System (CRS)

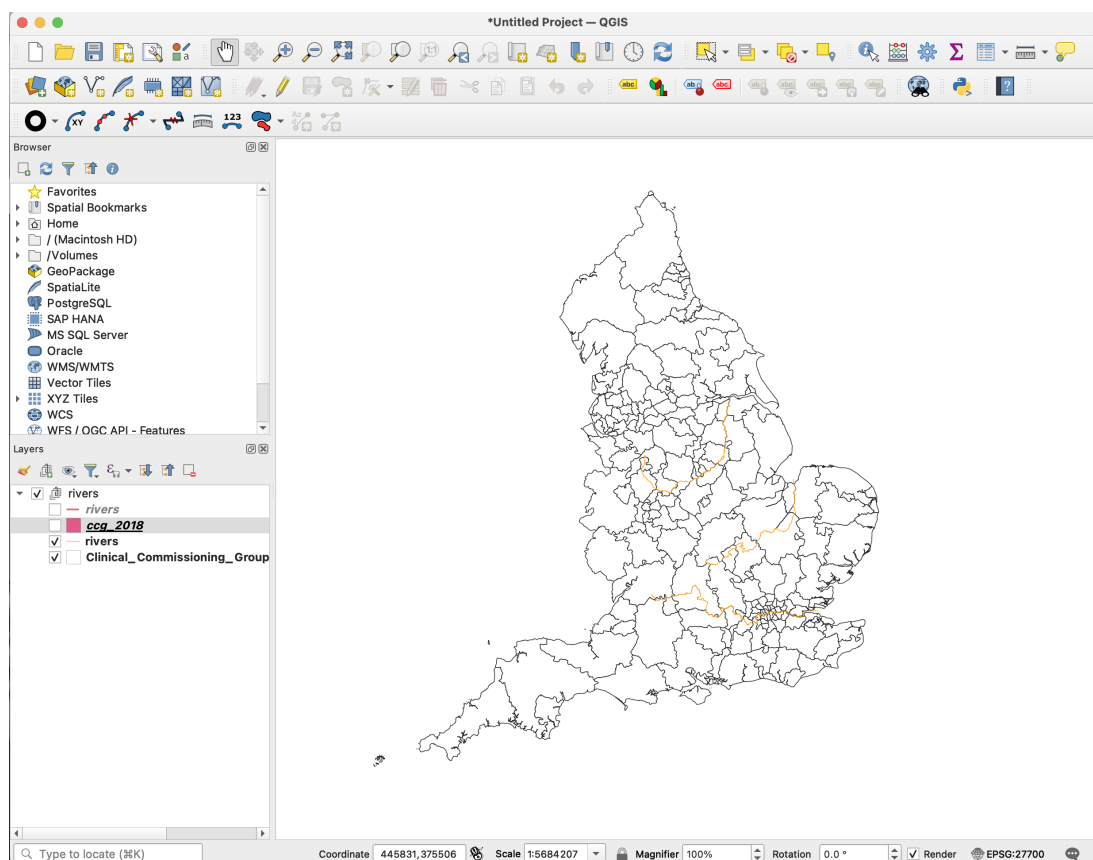


Figure 4.13: QGIS interface, showing the unified CRS (OSGB36) for both layers.

We then use QGIS to unify the CRS, and export both layers in GeoJSON format. See [Figure 4.14](#) and [Figure 4.15](#). The unified layer is shown in [Figure 4.13](#).

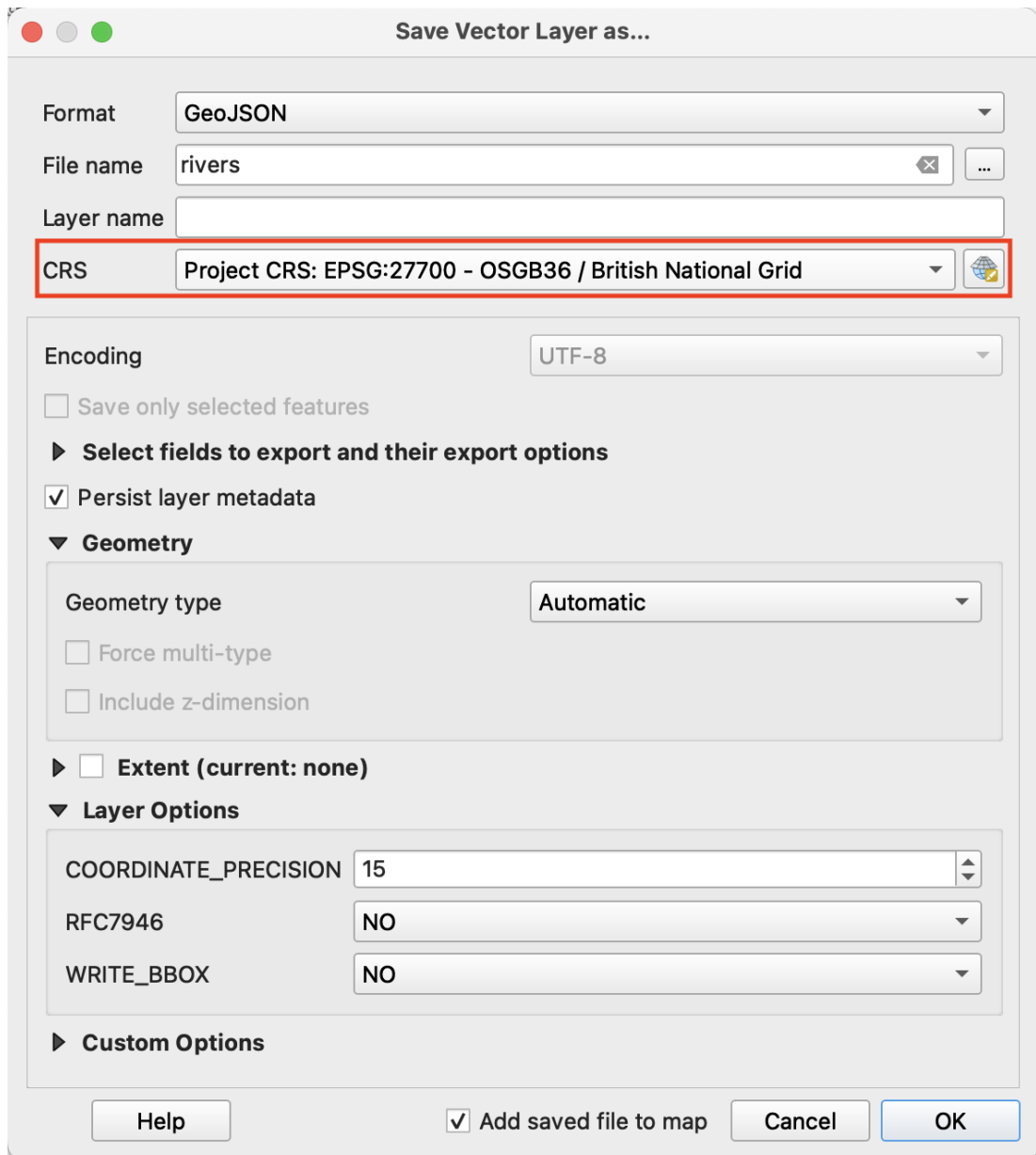


Figure 4.14: QGIS interface, exporting all rivers using the OSGB36 CRS in GeoJSON.

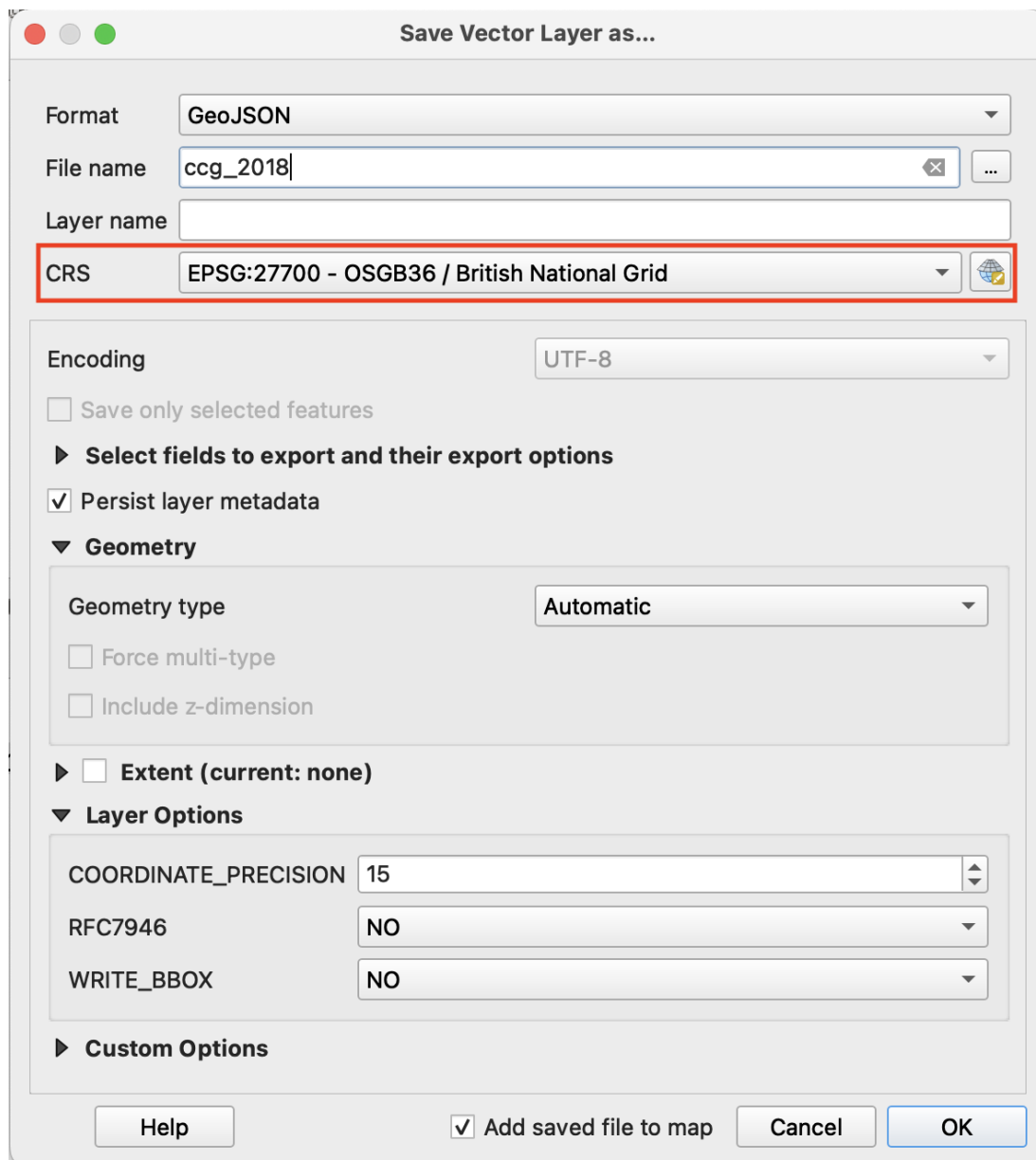


Figure 4.15: QGIS interface, exporting all NHS CCGs using the OSGB36 CRS in GeoJSON.

## Merge Shapefiles and Reduce File Size

We then merge two layers into one, and export it in the TopoJSON format using Mapshaper [324]. Mapshaper also supports the simplification of GeoJSON shapefiles. See Figure 4.16.

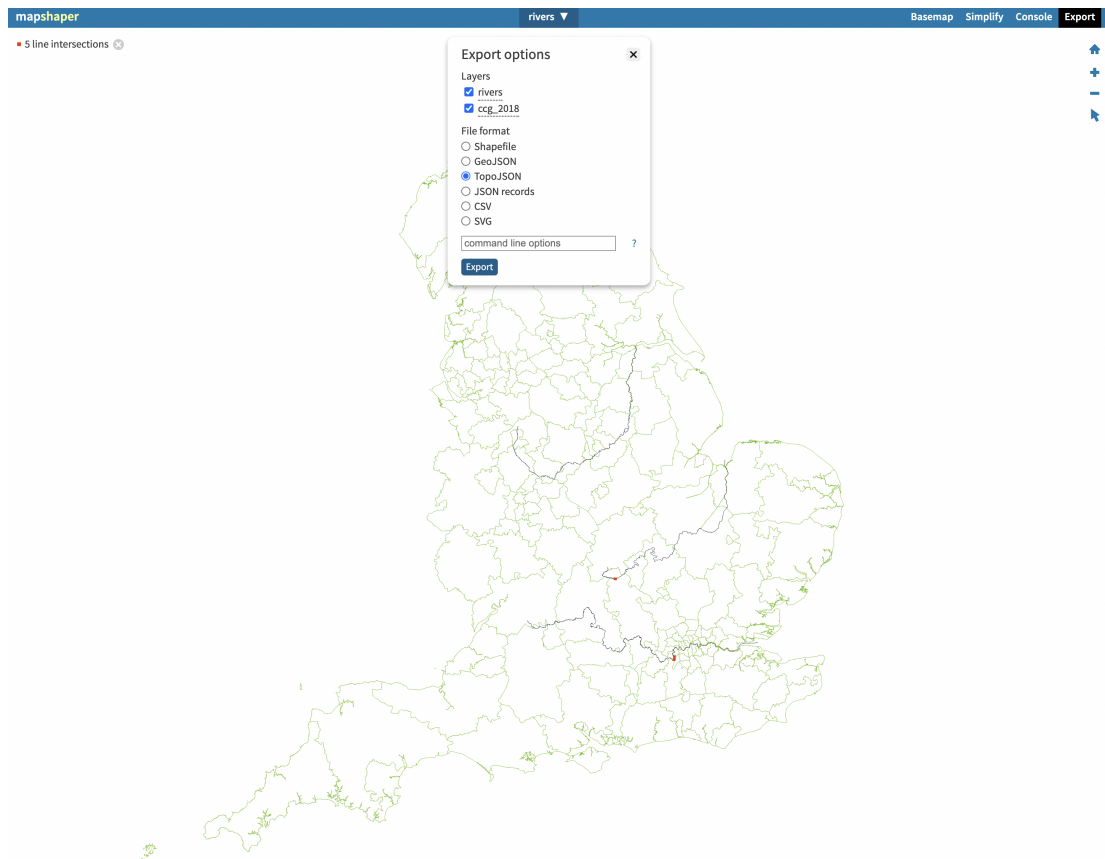


Figure 4.16: Mapshaper interface, merging all rivers with NHS CCGs into one layer, and export the merged layer in TopoJSON.

## 4.8.2 List of Likert Scale Questions

1. Including rivers in a cartogram is useful.
2. Including rivers increases the legibility of a cartogram.
3. Including rivers makes cartograms easier to understand.
4. Including rivers makes Clinical Commissioning Groups (CCGs) easier to locate on a cartogram.
5. Including rivers adds value to a standard cartogram.

Each of the five questions uses a 5-point Likert scale, where participants express their level of agreement with a statement about the role of rivers in cartograms, ranging from 1 (strongly disagree) to 5 (strongly agree).



## Chapter 5

### Time Series Map

Wang, Q., Bartolomeo, S. D., Dunne, C., Laramée, R. S., Litchfield, I., Weber, P., & Xu, K. (2024). Time Series Maps: Hierarchical Visualisation of Blood Glucose Time Series Data.

*“Knowledge and information are invisible. They have no natural form. It is up to the conveyor of the information and knowledge to provide shape, substance, and organisation.”*

– Donald A. Norman, the Scholar of User-Centred Design (1935 - present)

The chapter is based on our manuscript submitted to *EuroVis 2025*. In the previous chapter, we describe LetterVis, which explores the visualisation of unstructured text data. During LetterVis’ research and EHR Star’s presentation at EuroVis, we met health data experts interested in collaborating on visualising time series data, another commonly used structure for collecting EHR data. The need for a novel visual design that is both interactive and scalable to handle long time series data, inspired this chapter. The data sets we used in this chapter were all identified by our survey in [Chapter 2](#).

## Contents

---

<b>5.1</b>	<b>Introduction</b>	<b>143</b>
5.1.1	Terminology	144
5.1.2	Contributions	144
<b>5.2</b>	<b>Related Work</b>	<b>145</b>
5.2.1	Event Sequence Visualisation	145
5.2.2	Co-occurrence Pattern Visualisation	147
5.2.3	Clustering and Classification	147
5.2.4	Visualisation of Blood Glucose Data	148
<b>5.3</b>	<b>Event Specification and Extraction</b>	<b>149</b>
5.3.1	Natural Language Event Specification	149
5.3.2	From Natural Language to a Technical Event Specification	152
5.3.3	Event Extraction	154
<b>5.4</b>	<b>Time Series Map</b>	<b>155</b>
5.4.1	Hierarchy Construction	155
5.4.2	Time Series Map View	156
5.4.3	Color and Threshold	157
5.4.4	Other Views and Interactions	160
<b>5.5</b>	<b>Evaluation</b>	<b>161</b>
5.5.1	Data sets	162
5.5.2	Case Study	162
5.5.3	Health Data Experts Interviews	164
<b>5.6</b>	<b>Limitations and Future Work</b>	<b>171</b>
<b>5.7</b>	<b>Conclusions</b>	<b>171</b>

---



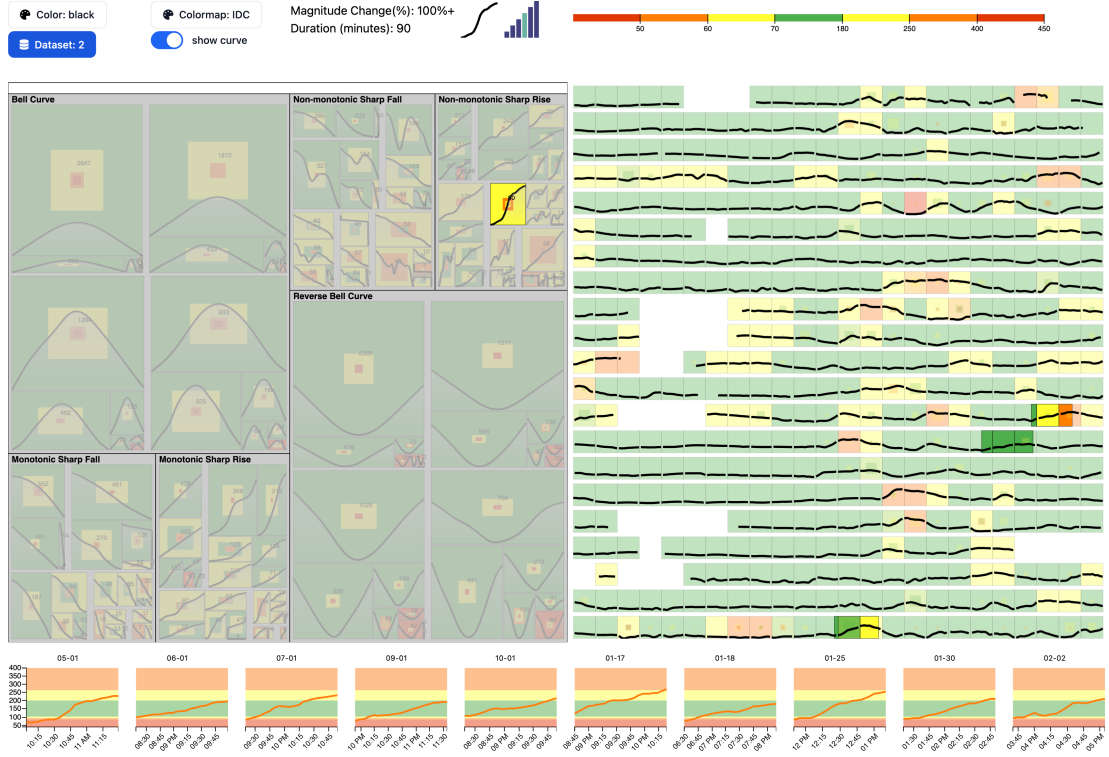


Figure 5.1: An overview of Time Series Maps to visualise blood glucose data. The hierarchical Time Series Map view on the left offers an overview of all the events, while the day-oriented view on the right provides a detailed view of the trends in blood glucose readings. The details-on-demand view at the bottom enables users to explore individual events and their corresponding blood glucose readings. In this figure, 8 months worth of data is rendered, which includes 68,000 events.

## 5.1 Introduction

The increasing prevalence of diabetes worldwide, coupled with the advancement in continuous glucose monitoring systems (CGM), has led to an exponential growth in the measurement of blood glucose readings [317]. This challenging data provides an invaluable opportunity to understand the nuanced relationship between behaviour and blood glucose fluctuations. Yet, the ability to effectively leverage this wealth of data is inherently dependent on our capacity to visualise and interpret it meaningfully.

Traditional methods for visualising time series data often rely on a linear mapping of the data over time on an axis [313]. While this approach is intuitive, it offers limited scalability, becoming unwieldy and challenging to explore when faced with data spanning over longer periods of time, e.g. days or months [313]. This constraint can pose barriers to our understanding of trends and patterns over longer periods [229]. It may hinder our ability to provide comprehensive and personalised care to individual

patients.

Event-based visualisation has been explored as an approach to visualise time series data. However, current event sequence visualisation designs face their own challenges. These techniques often struggle to manage a large number of events [101, 133], limiting the depth of insight that can be uncovered. Additionally, existing designs typically focus on categorical data [210, 231, 292, 264], overlooking the richness of continuous data such as blood glucose levels. Furthermore, these techniques often adopt generic approaches that may fail to capture the individual’s unique glycaemic responses, diminishing the visualization’s application for personalised care of patients with diabetes.

### 5.1.1 Terminology

In the context of this study, it is essential to delineate two types of data: time series and event-based.

Time series data is often regarded as a collection of observations recorded at regular intervals [119]. This type of data is crucial for analysing trends, patterns, or even forecasting future occurrences by examining the sequence of discrete data points over time.

On the other hand, event-based data consists of distinct observations of various types that are collected over time and organised in a sequence according to the particular entity to which the event is relevant [304]. Unlike time series data, which is marked by its regularity in time intervals, event-based data is characterised by its ability to capture sporadic, significant occurrences that provide insight into the dynamics of the observed entities.

### 5.1.2 Contributions

The Time Series Map is a novel hierarchical approach to time series data that combines advantages from both time series and event sequence visualisation. This hierarchical approach is carefully crafted to manage large-scale time series and event sequence data, offering a comprehensive overview and starting point for exploration and analysis. We then incorporate a case study to demonstrate the potential of our approach using two real-world CGM data sets. We develop and evaluate our approach with health data

experts to assess its effectiveness in gaining overviews of the data.

Our contributions include:

- A novel hierarchical visual design we call Time Series Map that offers an overview of time series data applicable to longer, arbitrary lengths of time,
- A general approach to identifying and extracting Event of Interest (EoI) from time series data,
- A scalable design to event-based visualisation that can handle a large number of events,
- Findings from usage scenarios and expert interviews in the diabetes treatment domain that demonstrate the utility of Time Series Maps for analysing blood glucose data.

The rest of this paper is organised as follows. [Section 5.2](#) discusses related work in time series and event-based visualisation. [Section 5.3](#) details our approach to extract events automatically from long time series data. [Section 5.4](#) presents our approach to visualising time series data using Time Series Maps. [Section 5.5](#) describes case studies designed to evaluate our work, and feedback provided by health data experts through interviews. We then discuss limitations and future work in [Section 5.6](#) and conclude the paper in [Section 5.7](#).

## 5.2 Related Work

We begin related work with surveys on time-oriented data visualisation [[86](#), [197](#), [223](#), [311](#), [304](#), [316](#)]. We then present related research work that aligns with our focus. To date, we have not found any related work that specifically focuses on providing overviews of long time series data. There is, however, research from multiple areas that focuses on the visualisation of time series data in general, which we consider here. We present prior work in these areas in the following sections.

### 5.2.1 Event Sequence Visualisation

Event sequence visualisation focuses on a type of temporal data consisting of a series of incidents that unfold over time [[311](#)]. Visualisation techniques are essential for reducing

Work	Visual Representation	No. of Events Rendered (n)	Subject of Data (UMLS Code)
Monroe et al. [122]	Scatterplot, Node-link	$n < 3,000$	Pharmacology (C0031330)
Gotz et al. [132]	Scatterplot, Node-link	$n < 50$	Medical history (C0262926)
Malik et al. [152]	Glyphs, Timeline	$n_c < 100$	Care of intensive care unit patient (C0010337) Patient care management (C0030677)
Perer et al. [154]	Bubble chart, Sankey diagram	$n_c < 100$	Disease progression(C0242656)
Kwon et al. [170]	Glyphs, Histogram, Node-link	$n < 1,000$	Patient timeline (C1705821)
Guo et al. [210]	Glyphs, Histogram, Node-link	$n < 1,000$	Chronic obstructive airway disease (C0024117)
Guo et al. [231]	Hierarchy-based, Timeline-based	$n_c < 500$	Care of intensive care unit patient (C0010337)
Zhang et al. [251]	Glyphs, Timeline-based, Violin plot	$n \approx 200$	Diabetes mellitus, insulin-dependent (C0011854)
Gotz et al. [261]	Chart-based, Hierarchy-based, Timeline-based	$n \approx 6,000$	Opioid-related disorders (C0027412)
Jin et al. [264]	Node-link	$n < 100$	Cardiovascular diseases (C0007222) Respiratory tract diseases (C0035242)
Di Bartolomeo et al. [290]	Sankey-based, Timeline-based	$n \approx 4,000$	Diabetes mellitus, insulin-dependent (C0011854)
Jin et al. [292]	Glyphs, Matrices, Node-link	$n_c < 200$	Pneumonia (C0032285)
Guo et al. [303]	Bar Charts, Histograms, Node-link	$n < 500$	Care of intensive care unit patient (C0010337)
Magallanes et al. [308]	Histogram, Matrix, Stacked Bars, Tree	$n_c < 1,500$	Atrial Fibrillation (C0004238)
Time Series Maps	Treemap, Line Charts, Histogram	$n > 68,000$	Blood glucose management (C1638311)

Table 5.1: The table presents a list of event sequence visualisation research with the scope of EHR published over the last decade. We categorise the papers based on their *Visual Representation*, *Number of Events Rendered*, and *Subject of Data (UMLS Code)*, where  $n_c$  denotes the number of groups and clusters rendered.

the visual complexity of events and to facilitate the identification of patterns and trends.

Here, we focus on the visualisation of event sequences in the context of EHR. In Table 5.1, we adopt the taxonomy of event sequence visualisation by Guo et al. [304], and categorise the previous work using *Visual Representation* (visualisation techniques used to render event sequences) and *Data Scale* (data granularity). In addition, we also include a *Subject of Data* column to utilise the Unified Medical Language System (UMLS) [32] to provide a brief description of the application in each work. Our work also renders a data set with 68,000 events collected over 8 months easily.

Compared to the work listed in Table 5.1, our work focuses on the visualisation of sequence collections of long diabetes data sets, featuring a hierarchy-based visual representation as an overview. It incorporates linked chart-based and timeline-based views to support user interactions and detail-on-demand views. Our approach is designed to be scalable and capable of accommodating significantly larger data sets,  $n > 68,000$ .

### 5.2.2 Co-occurrence Pattern Visualisation

Events extracted from time series data at different locations form a co-occurrence pattern, which can be visualised to explore temporal relationships between events occurring at two or more locations [240]. Visualisation of co-occurrence patterns is a well-studied area, with research focusing on the visualisation of co-occurrence patterns in the context of human mobility [183, 237, 312], which often establishes an association between spatial and temporal aspects of the data, such as GeoChron [314] that presents spatial-temporal patterns from large-scale time series data. Often, the focus of these works is pattern recognition and event extraction before establishing temporal relationships between events, which is not the central theme of our work. Our work is focused on providing an overview of long-time series data as a starting point for exploration.

### 5.2.3 Clustering and Classification

Utilising machine learning techniques has become a popular approach to expedite the processes of clustering and classifying time series data, in order to address the scalability challenges intrinsic to manual processing.

Ali et al. [223] classified clustering techniques applied to time series data into three

categories: *whole time series clustering*, *subsequence clustering*, and *temporal proximity and value clustering*. While we did not incorporate a machine learning algorithm, our approach is similar to subsequence clustering, which leverages a sliding window to extract and classify events from the time series data based on the event categories described in [Section 5.3](#).

## 5.2.4 Visualisation of Blood Glucose Data

IDMVis [\[251\]](#) is a novel visualisation tool designed to enhance clinical decision-making in Type 1 diabetes management. It addresses the challenge of visually integrating diverse data sets, such as manual logs and medical device data, into a cohesive visual representation. IDMVis includes features such as the folding and aligning of records around key events, dynamic timeline scaling, and statistical summarisation. A qualitative evaluation was conducted with six clinicians, underscoring IDMVis’ potential to transform data interpretation and decision-making processes through advanced visualisation techniques. We use data from this paper in our evaluation.

Di Bartolomeo et al. [\[290\]](#) present Sequence Braiding, a novel visualisation technique for the overview analysis of temporal event sequences and attributes, motivated by blood glucose data visualisation. The technique uses a layered directed acyclic network, aligning temporal events and attribute groups simultaneously. The paper’s central focus is on the development of an N-layer network layout algorithm, emphasising rank assignment and intersection reduction for optimal sequence alignment. A case study on type 1 diabetes treatment is used to demonstrate the technique’s application, highlighting its effectiveness in assisting users to quickly understand patterns and trends in complex temporal event sequence data.

Marjorie [\[315\]](#) is a visual analytics tool tailored for Type 1 diabetes, enhancing data analysis in clinical consultations. It employs modified horizon graphs for blood glucose visualisation, and integrates hierarchical clustering to present insulin and carbohydrate data. Marjorie features semantic zooming for detailed data exploration and utilises dynamic time warping to identify specific glucose patterns. Marjorie is validated through feedback from the diabetologist and a real patient data case study, offering critical insights for effective diabetes management, focusing on clear data interpretation and

pattern recognition in clinical settings. We also use the data featured in this paper as part of our evaluation, as well as validation through feedback from health data experts.

Our work resembles the above work, as we focus on the visualisation of blood glucose data. However, our work differs in that we offer a novel hierarchy-based visual representation as an overview, enabling users to explore the entire data set by starting with common patterns and outliers highlighted in the overview. We also adopt different approaches to identify, search for, and extract events that merit attention within the data.

## 5.3 Event Specification and Extraction

Rather than focus on the entire time series data, we start our focus with EoI. EoIs will change depending on the application. In our case, we focus on blood glucose EoIs. However, the EoIs we identify and the process we describe are very generic and will apply to any time series data, as can be seen in [Figure 5.2](#).

We begin with the event specification, where we interview a health data expert to identify EoIs and event categories. Second, we transition the expert’s natural language specification to a technical specification more amenable to automatic EoI search and identification. Third, we describe an automatic approach to extracting EoIs from time series data. We then propose a visual hierarchy to represent those categories and their events, which are then used to generate overviews using a Time Series Map. The choice of colormaps is then described. Finally, we describe how we link the event overviews to their respective day-oriented views, and how we enable exploration with multiple-linked views and details on-demand.

### 5.3.1 Natural Language Event Specification

As patients with type 1 diabetes require regular insulin injections or an insulin pump to manage their blood glucose levels, a CGM is often used to continuously monitor the patient’s blood glucose levels throughout the day. Through an interview with a health data expert, we obtained a list of important event categories initially expressed in natural language. Some events are especially noteworthy, as they serve as indicators

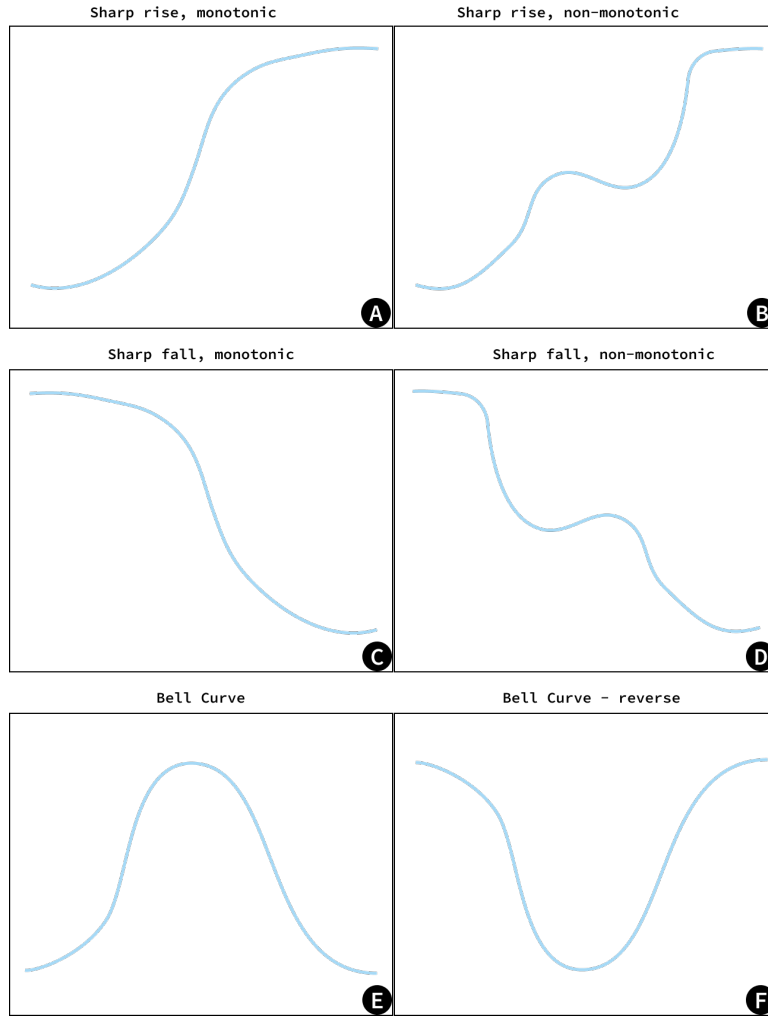


Figure 5.2: Six shape-based event categories reflecting blood glucose levels that may require attention. Category (A) depicts a monotonic sharp rise, while Category (B) illustrates a nonmonotonic sharp rise. Categories (C) and (D) represent monotonic and nonmonotonic sharp falls, respectively. Category (E) is characterised by a bell curve. Category (F) depicts a reverse bell curve.

of blood glucose levels that may require attention. We initially divide these events into two top-level categories: *threshold-based* and *shape-based*.

**Threshold-based Events:** Threshold-based events are simply times when blood glucose readings pass a given level of importance. Drawing from the glucose target ranges and their respective categories as shown in the diabetes research literature [117], see Figure 5.9, we use these ranges to identify important threshold-based events.

**Shape-based Events:** Shape-based events are important periods of fluctuating blood glucose initially indicated by a health data expert using a hand-drawn sketch. See Figure 5.2 for a re-creation of the hand-drawn sketches provided by a health data expert. These events indicate that the glucose reading profile exhibits a given shape,



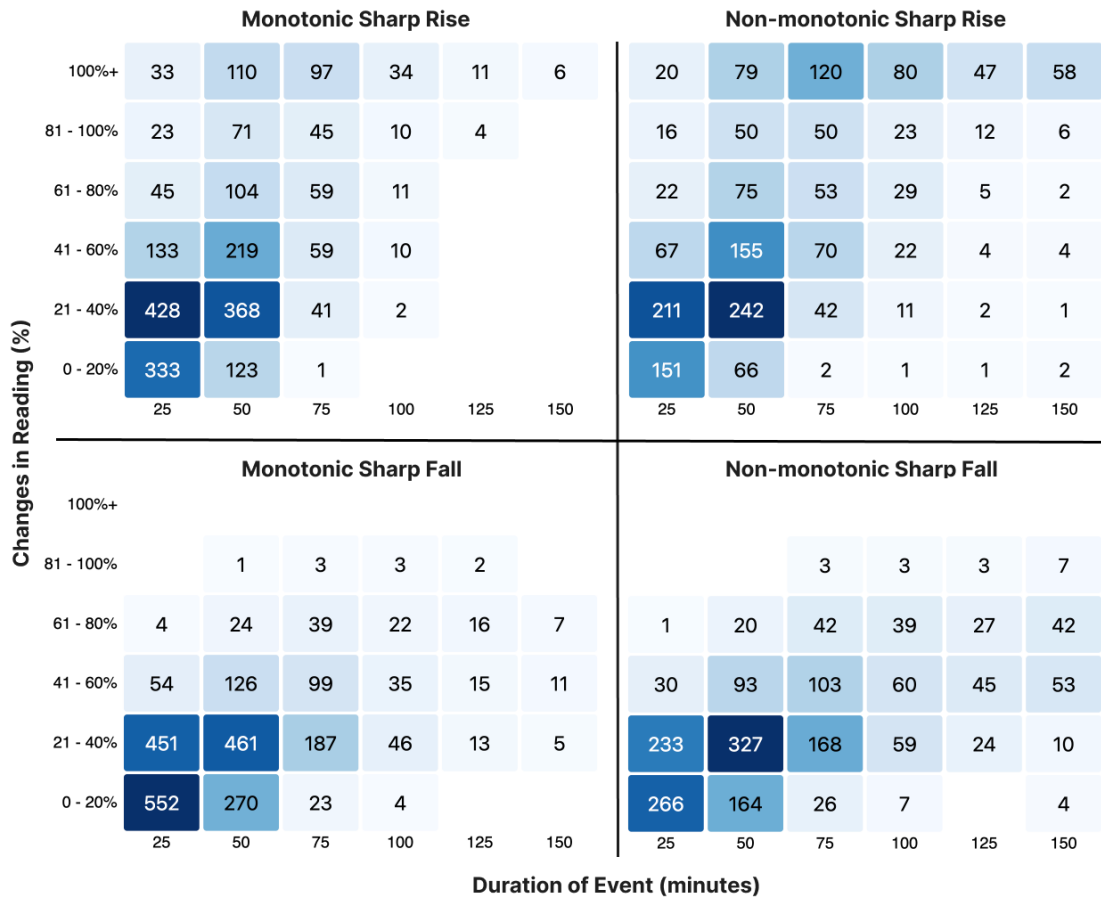


Figure 5.3: This pair of 2D histogram matrices depict the distribution of both monotonic and nonmonotonic sharp rises and falls captured automatically using a duration indicated by the x-axis. The y-axis represents the change in readings as a percentage, and the x-axis represents the duration of events in minutes. Colour is simply mapped to number of events in each histogram bin. In this case, 8,513 events are automatically identified from the Marjorie data set [315].

such as glycaemic variability, characterised by spikes or drops. Such events may be triggered by inconsistent carbohydrate intake, incorrect insulin doses, or poor adherence to medication.

**Sharp Changes:** Indicate that the glucose reading exhibits a sharp rise. We divide sharp rises into two subcategories: monotonic and nonmonotonic. Specifically, a monotonic sharp rise strictly follows an increasing trajectory. See Figure 5.2 (A). In contrast, a nonmonotonic sharp rise takes into account readings that do not strictly increase successively. See Figure 5.2 (B). This consideration is based on the potential for the readings to exhibit fluctuations with a sharp increase. We treat sharp falls similarly.

**Bell Curves:** Indicate that the glucose reading exhibits a Gaussian curve shape.

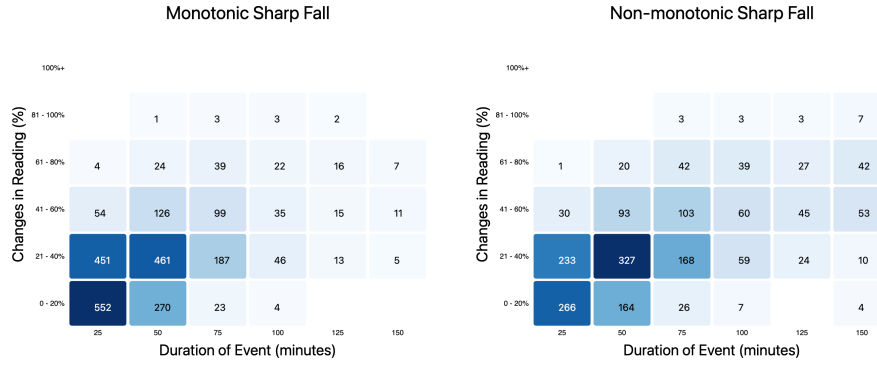


Figure 5.4: Similar to Figure 5.3, this 2D histogram matrix depicts both the number of monotonic and nonmonotonic sharp falls captured automatically.

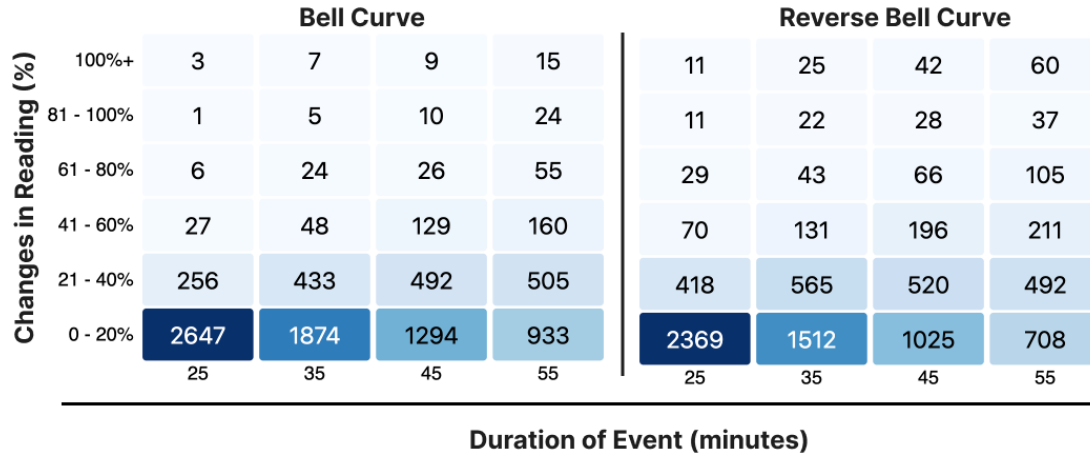


Figure 5.5: A sample 2D histogram matrix depicts bell curve and reverse bell curve EoIs using a sliding window ( $k = 12$ ). The y-axis represents the percent change in glucose readings, calculated using the minimum and maximum values in each captured curve, while the x-axis is mapped to event duration. The distribution reveals that the subject's glucose readings show frequent but small spikes and drops, which could be used to infer the stability or variability of glucose levels over time. The majority of the events fall within the 0 - 20% change bracket. 17,679 curves are captured and rendered from the Marjorie data set [315].

See Figure 5.2 (E) and (F) for an illustration. This may also indicate a sharp rise followed by a sharp fall and vice versa. We capture bell curves, as they can potentially indicate high glycaemic variability, which contributes to multiple complications related to diabetes and has a negative impact on the quality of life of a patient [155, 307].

### 5.3.2 From Natural Language to a Technical Event Specification

Starting from the natural language description plus some initial sketches of important EoIs, we set out to turn this into a technical specification of events in the time series

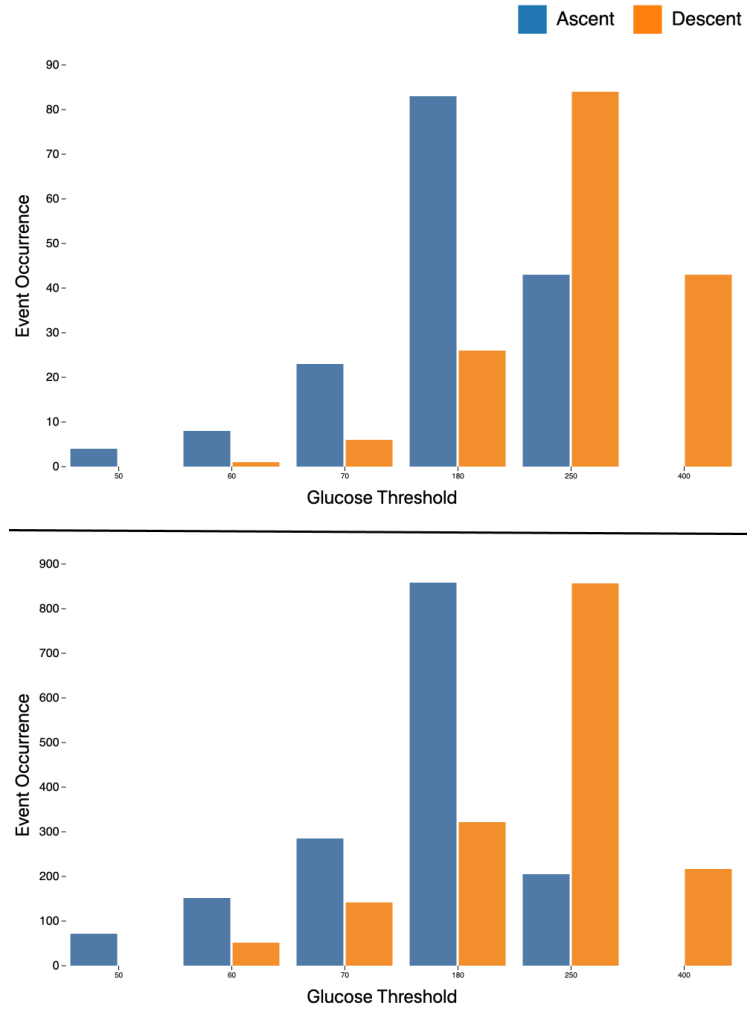


Figure 5.6: This grouped bar chart depicts the distribution of threshold events in the Marjorie data set [315]. The y-axis represents the number of threshold events, and the x-axis represents the blood glucose thresholds. Ascent threshold events, e.g. 49 to 50, are shown in orange, and descent threshold events, e.g. 50 to 49, are shown in blue.

space, such that we can automatically search for these patterns.

**Threshold-based Events:** For threshold-based events, we simply identify the times at which blood glucose levels cross the important thresholds presented in Figure 5.9. A sample distribution of threshold-based events is shown in Figure 5.6.

**Shape-based Events:** To transform the natural language specification into a technical specification, we created digital versions of the curves based on synthetic data manually. For shape rises, we started with four data points representing a sharp increase. A sharp rise is considered to be an increase of 20 mg/dL over a period of 20 minutes ( $k = 4$ ). We then added to the duration of increments up to 150 minutes. For identifying bell curves, we begin by creating synthetic curves with perfect symmetry to

resemble the sketches, similar to those illustrated in [Figure 5.2](#). The lengths of these synthetic curves are set to 6, 8, 10, and 12 sample data points, which correspond to 30, 40, 50, and 60 minutes, respectively.

### 5.3.3 Event Extraction

After transforming the natural language description of EoIs to a more technical specification, we can then use automatic EoI identification and extraction methods. We then apply STUMPY [\[239\]](#) with the synthetic curves to capture individual curve-shaped events in the data automatically.

**Extraction Library:** Siebert et al. [\[296\]](#) conduct a comprehensive systematic review focusing on time series analysis packages available in Python. The authors analyse and categorise a total of 40 packages, considering various factors such as the analytical tasks implemented, data preparation methods, and the means of evaluating the results. Drawing on their insightful findings, we carefully select packages that align with our specific requirements. Among these choices, we opt to utilise STUMPY [\[239\]](#), a popular Python library for its efficient computation of the matrix profile.

**Matrix Profile:** To improve the scalability of our work, we use matrix profile computations to effectively categorise and group events [\[239\]](#). At its core, the matrix profile is constructed by computing the Euclidean distance between subsequences of a time series and their nearest neighbours [\[184, 185\]](#), making it an effective technique to reveal patterns and aid in anomaly detection. This computation involves calculating the distances between subsequences and their neighbours. The matrix profile also supports various tasks such as motif discovery, thus improving decision-making and our understanding of time series data [\[221, 234\]](#).

**Extraction Analysis:** We utilise 2D histogram matrices to illustrate the sample distribution of EoIs in the Marjorie data set [\[315\]](#). [Figure 5.3](#) shows the distribution of sharp rises and falls, both monotonic and nonmonotonic. The y-axis represents the percentage change in readings, while the x-axis represents the duration of EoIs in minutes. For these 2D histograms, we use a continuous blue hue colormap from ColorBrewer [\[28\]](#), where the colour is simply mapped to the number of EoIs in each duration-magnitude bin. We plot 2D histograms in order to obtain an understanding

and overview of the distribution of EoIs in both duration and magnitude for each type of curve in [Figure 5.2](#). The data reveals that sharp falls are typically short-lived with minor changes in readings, while sharp rises tend to be more prolonged and involve larger changes. Nonmonotonic EoIs, compared to monotonic ones, are more persistent and exhibit larger changes.

Another 2D histogram matrix depicts the bell curve and reverse bell curve EoIs in the Marjorie data set [\[315\]](#). See [Figure 5.5](#). The distribution reveals that the patient’s glucose readings show frequent but small spikes and drops, which could be used to infer the stability or variability of glucose levels over time.

## 5.4 Time Series Map

In this section, we present our scalable design for Time Series Maps, specifically tailored to render visual overviews of large-scale time series data, such as blood glucose readings. This design is informed by feedback from health data experts to validate its relevance and effectiveness in real-world applications. We begin by detailing the construction of the hierarchical structure that underpins our approach, followed by an in-depth discussion of the design elements and colormaps employed to enhance data visualisation. Finally, we conclude this section by exploring additional views and interactive features that complement and enrich the exploration of the rendered data sets.

We provide a demonstration video to illustrate the design, available at <https://youtu.be/TnlyZDQCpQE>. There is a live demo of Time Series Maps at <https://tsm.wangqiru.com/>, with all supplementary materials available at <https://doi.org/10.17605/OSF.IO/7B6VW>.

### 5.4.1 Hierarchy Construction

At the core of our approach is to build a scalable hierarchy of EoIs. Our hierarchy is derived from the unique categories of the six shape-based event groups in [Figure 5.2](#). The hierarchy is composed of 5 levels: 1) Event at the top, 2) Sharp rises, falls, and bell curves underneath, 3) monotonic and nonmonotonic, 4) the histogram bins described in [Section 5.3.2](#) and, 5) the individual EoIs at the bottom. See [Figure 5.7](#) for an

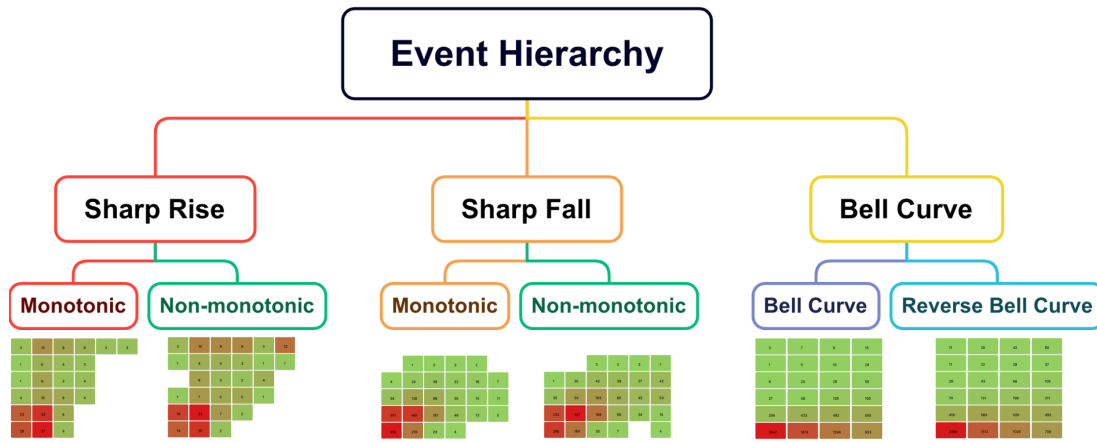


Figure 5.7: The hierarchy is derived from the unique shapes of the six shape-based event groups presented in Figure 5.2. Each group is further subdivided into a range of durations as indicated in Figure 5.8, which are mapped to the x-axis, while the changes in reading are mapped to the y-axis.

illustration. In the Time Series Map, each group is represented by a concentric rectangle that illustrates the distribution of glucose profile ranges within the group itself.

## 5.4.2 Time Series Map View

A treemap is an efficient space-filling visualisation technique capable of depicting hierarchical structure and enabling intuitive comparisons among categories [6, 7]. We adapted a treemap to provide an overview for a number of reasons: 1) a treemap is scalable, 2) a treemap is a well-known visual representation, and 3) a treemap is not bound to a linear mapping of time to an x-axis like most other time-oriented representations. The treemap shown in Figure 5.8 is constructed using the hierarchy illustrated in Figure 5.7. The size of each parent node is determined by the number of children in the corresponding subcategory. Each node also contains a set of concentric rectangles that depicts the frequency and distribution of threshold events in that group (Figure 5.11).

Users can toggle the visibility of the curve in the Time Series Map view and choose the curve's color. This user option was introduced as a feature based on a health data expert's suggestion that the curve obscured the node, and removing it would enhance clarity.

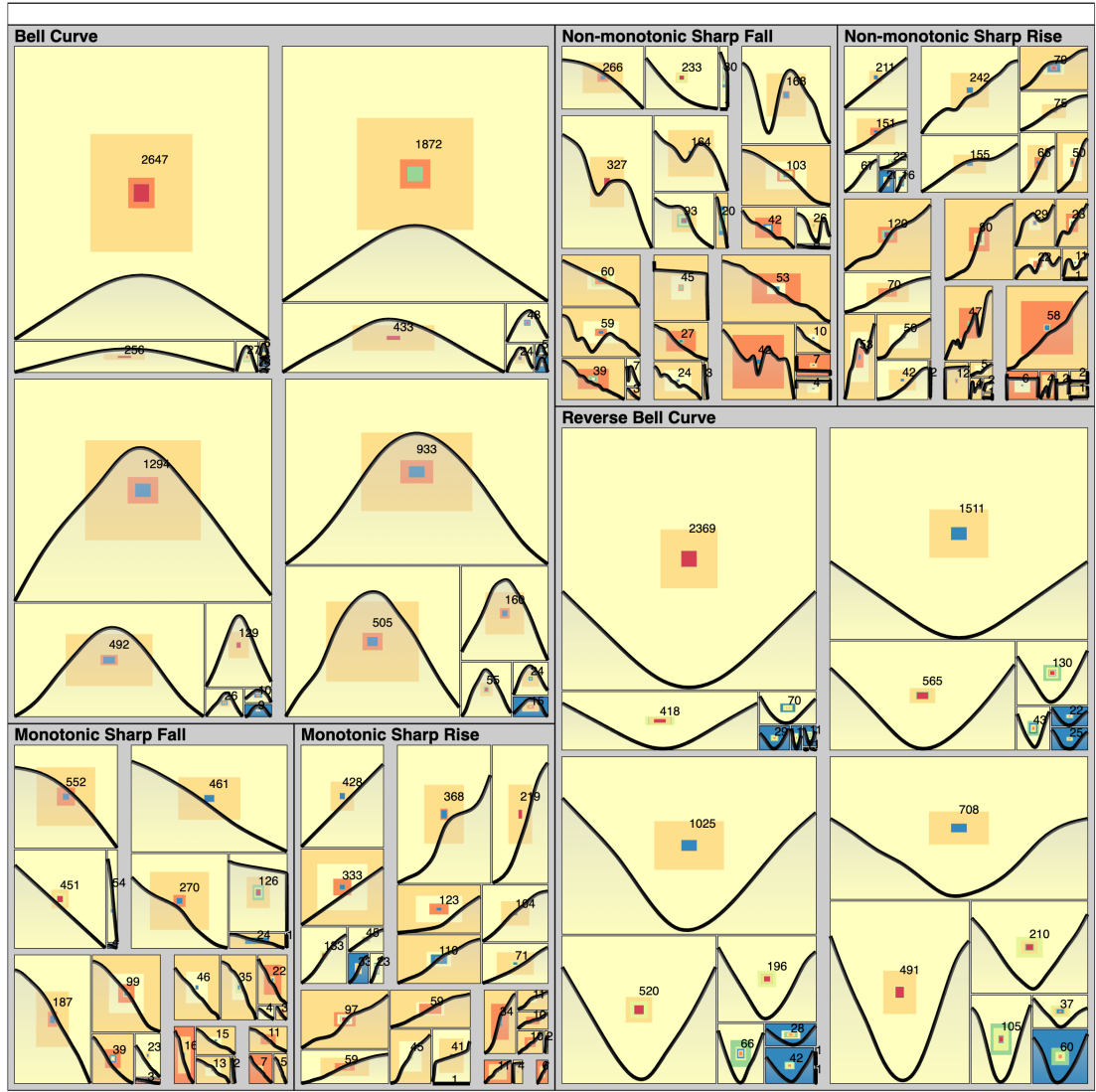


Figure 5.8: A Time Series Map constructed based on the hierarchy shown in Figure 5.7. It provides an overview of all the events, with the size of each parent node determined by the number of children in the corresponding event category. Each child node also contains a set of concentric rectangles that depict the frequency and distribution of threshold EoIs for that group. Superimposed over each child node is a curve provides a representative trend for the category, illustrating each group in Figure 5.2. This view enables quick identification of both the volume and characteristics of EoIs within each category. The colour scale is shown in Figure 5.10 bottom.

### 5.4.3 Color and Threshold

We experimented with a number of colour mapping options and arrived at two guided by our interviews with health data experts. The first and default colour scheme is a standard categorical colormap taken from the International Diabetes Center, as shown in Figure 5.9. In this scheme, four distinct colours are used to represent the glucose target ranges and their respective categories: 1) **Dangerously High/Low**; 2) **Very High/Low**;

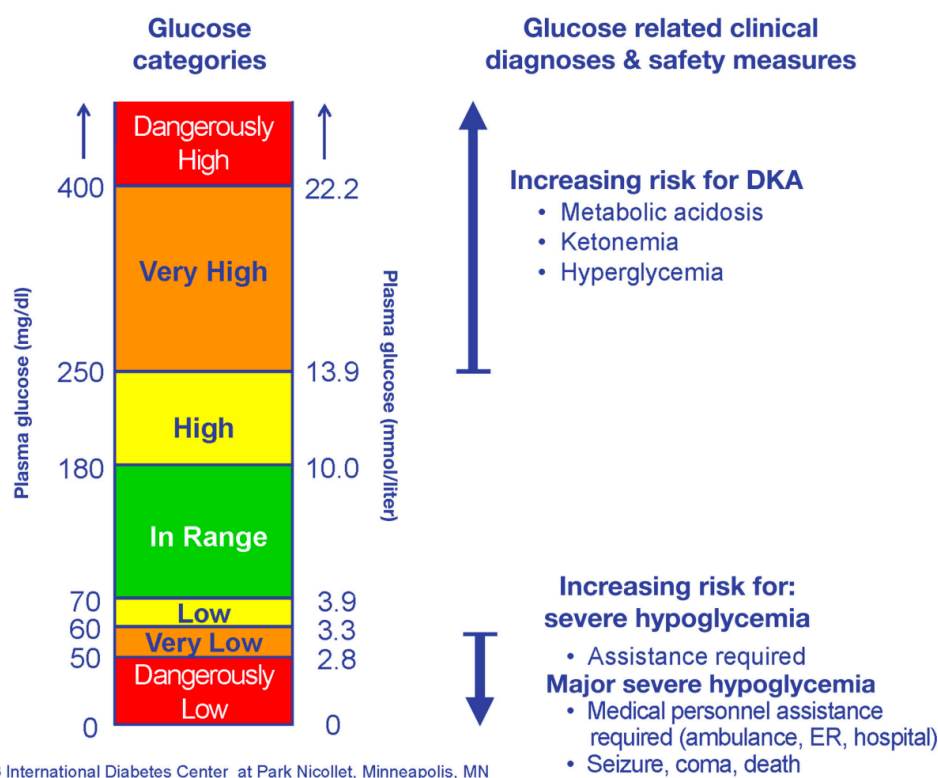


Figure 5.9: Important glucose target ranges, thresholds, and categories, proposed by the International Diabetes Center [117]. We use this as one of our colormap options, due to the importance of threshold-based events as guided by the health data experts we worked with. Figure reproduced from Bergenstal et al. [117].

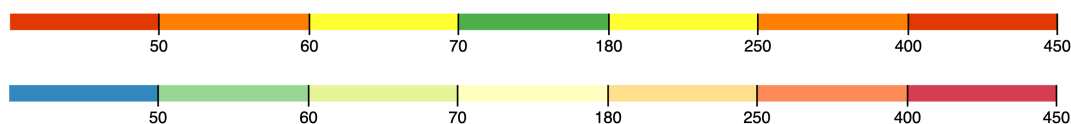


Figure 5.10: Two colormaps are available for representing glucose reading categories. Top: a colormap from the International Diabetes Center, as shown in Figure 5.9. Bottom: a sequential colormap derived from ColorBrewer [28], as suggested by a health data expert.

3) **High/Low**, and 4) **In Range**.

Guided by the feedback of the health data expert, we incorporate a second colour scheme, a sequential colormap with a 7-category colormap derived from ColorBrewer [28], to make a distinction between the low and high categories because, as observed by one expert, the low blood glucose values may result from different circumstances than the high values. See Figure 5.10.

We experimented with a number of different options for depicting the threshold events. Since they are not exclusive to curve-based events, we integrate them directly into the treemap layout. We considered different options including inserting a histogram



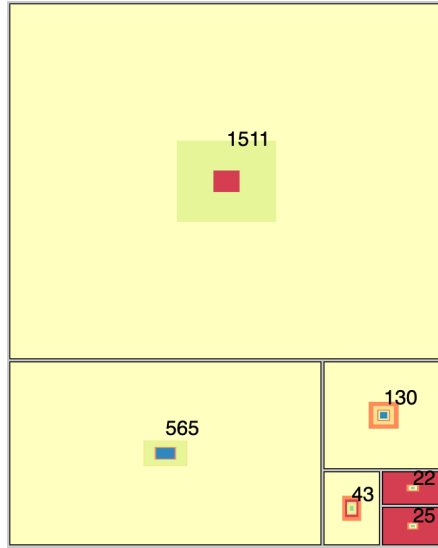


Figure 5.11: The figure illustrates treemap nodes with a concentric square colour design. Each node’s colours are determined by the distribution of glucose readings per threshold category represented by the node. The numbers indicate the total number of EoIs for the child node.

inside each treemap rectangle, to indicate the different instances of threshold events in each category of shape-based events. However, this visual design is not optimal according to Tong et al. [201]. We also experimented with mapping the colour of each treemap node to the average blood glucose value amongst its children. However, we found that dangerously high and dangerously low events were obscured by this choice. Thus, we wanted a colour design that captures each category clearly and that integrates easily into the Time Series Map design.

We chose a concentric square colour design seen in both the Time Series Map and day-oriented views, based on the existing colormaps. See Figure 5.11 for an example of the concentric design. The concentric design represents the number of events that cross the blood glucose threshold per glucose range depicted in Figure 5.9 and within each set of child EoIs. The size and rendering order of the concentric colours within each rectangle are determined by the distribution of blood glucose readings, with the outermost shape representing the most frequent threshold events. The smallest, innermost square represents the least frequent range of blood glucose readings. By rendering the least frequent range of blood glucose readings on top, they are not obscured by these larger, more frequent categories.

#### 5.4.4 Other Views and Interactions

Here, we describe the additional views employed to complement and enrich the exploration of the rendered data sets.

##### Day-oriented View

The day-oriented view ([Figure 5.12](#)) is critical because it maps time to the more traditional x-axis to facilitate the interpretation of the Time Series Map view. It is also the view with which most users are familiar. Each row in the day-oriented view represents a day (24 hours). Each day is divided into hourly blocks by default from midnight to midnight. Aligning days on top of one another supports observation of daily patterns. This choice was supported by one of the health data experts we interviewed. Because not all days are visible in this view, the user can scroll up and down to observe all the rows showing the remaining days.

##### Details-on-demand View

The details-on-demand view enables in-depth analysis of individual EoIs through juxtaposed line charts, an example of which is shown in [Figure 5.1](#) bottom. As the view with the finest granularity, it provides close observation of glucose readings and facilitates the detection of patterns in glucose readings. The view is populated with charts by selecting a specific rectangle in the Time Series Map view.

##### Interaction and Exploration

Utilising three views, our implementation supports linked interaction to facilitate exploration. The Time Series Map view, as an overview, provides the entrance to the exploration. Selecting a specific node mapped to an event category triggers updates in both day-oriented and details-on-demand views to display the child EoIs. For an illustrative example, refer to [Figure 5.1](#) and the supplementary video. By selecting a child node in the Time Series Map view, all events within that group are highlighted in the day-oriented view, displayed with the time and duration of each event. The details-on-demand view is updated with an individual EoIs, each of which is a member of the selected group. Clicking on a line chart will navigate the user to the corresponding

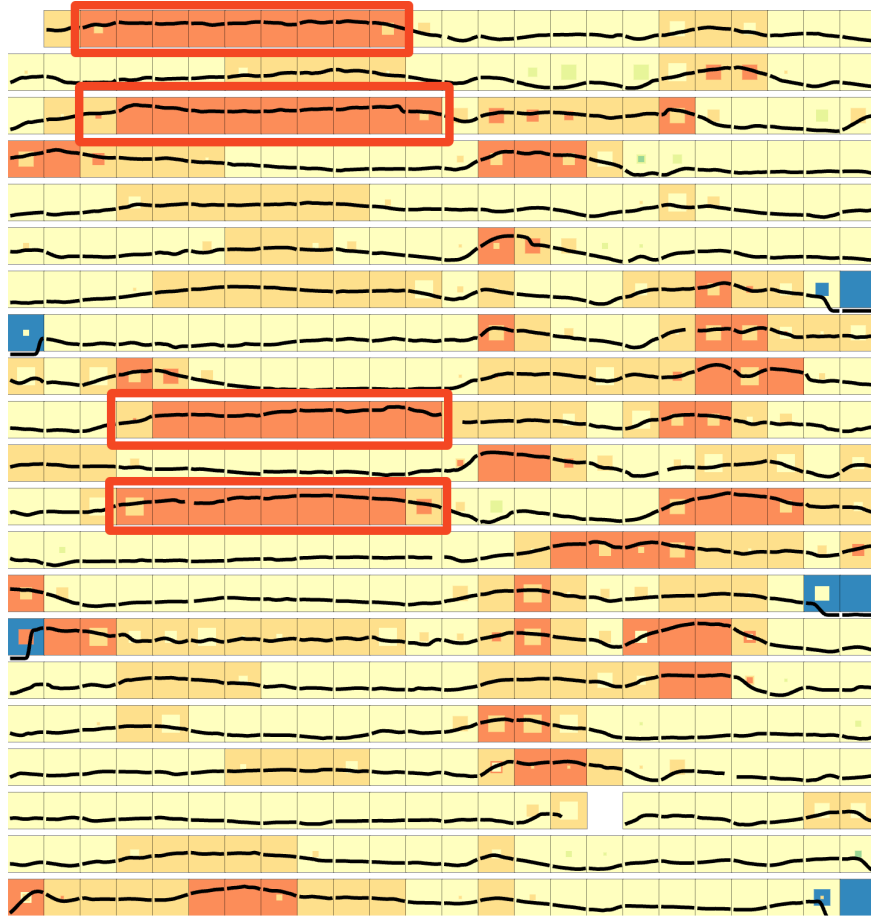


Figure 5.12: The day-oriented view provides a more traditional depiction of blood glucose readings with each row representing one day divided into 24 hours. The x-axis represents the time of day, while the y-axis of each row represents the blood glucose readings in mg/dL. Each colored rectangle corresponds to one hour, with the horizontal curve within indicating the blood glucose readings. The gaps in the graph, where the rectangles are absent, are periods when readings are missing, indicating times when the continuous glucose monitor did not record. The red outline rectangles highlight patterns observed in case study 2.

event in the day-oriented view for further examination.

## 5.5 Evaluation

In this section, we describe two data sets we use throughout our study. We then detail two evaluation methods used in our study. We first conduct case studies to demonstrate the effectiveness of Time Series Maps. The case studies in [Section 5.5.2](#) provide a contextual and nuanced exploration of how Time Series Maps function with a real-world time series data set. We then invite health data experts for guided interviews, as described in [Section 5.5.3](#). Experts share their specialised knowledge and experience

and offer valuable insight into the effectiveness and appropriateness of Time Series Maps.

### 5.5.1 Data sets

We use two open data sets of continuous blood glucose readings. The first data set is obtained from IDMVis [251], which contains a CSV file of blood glucose readings measured from two CGMs (Tidepool and Nightscout) collected from a patient with type 1 diabetes over a period of 26 days. The CGMs measure glucose readings every 5 minutes. The data set features 1) Time - the time of the reading, e.g., 2017-08-23 00:03:52.591; 2) Source - the source of the reading, either from Nightscout or Tidepool (CGM). We only use readings from Nightscout, as we observe that the readings from Tidepool contain irregularities and missing data; 3) Unit - the unit of the reading, milligrams per deciliter (mg/dL); and 4) Glucose reading - the glucose reading, e.g., 152.6.

The second data set was obtained from Marjorie [315], which contains a CSV file of blood glucose readings measured from a patient with type 1 diabetes over a period of 8 months. The CGMs measure glucose readings every 5 minutes. The data set features: 1) Day - the date of the reading, e.g., 23.12.2021; 2) Time - the time of the reading, e.g., 22:12; and 3) Glucose reading - the glucose reading, e.g., 152.

For the implementation and evaluation of our approach, we use both data sets. Given that the second data set is longer, we use it to demonstrate the scalability of our technique. Figures presented in this paper are derived from the second data set. However, for the case study we use the first data set because it features supplementary data such as a diary of food intake and insulin dosage.

### 5.5.2 Case Study

In this section, we explore case studies that demonstrate the effectiveness of Time Series Maps.

All case studies are included in our video, available at <https://youtu.be/TnlyZDQCpQE>.

**Case Study 1: Finding and Analysing Unusual Rises:** The Time Series Maps

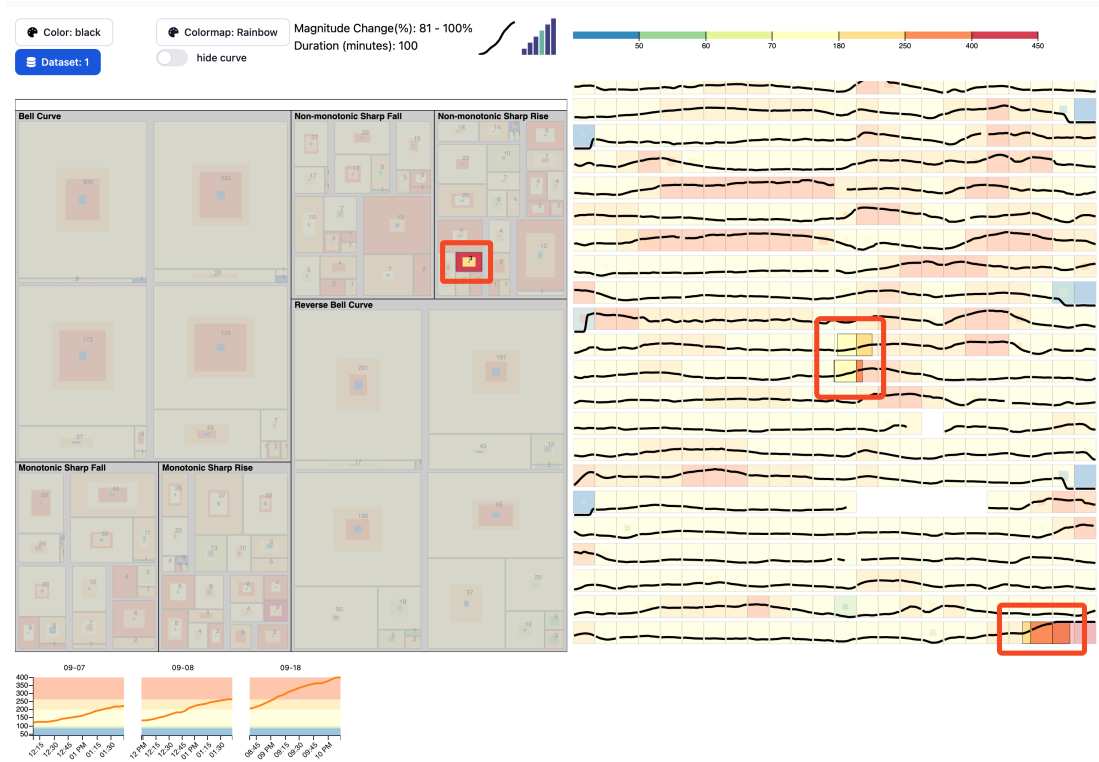


Figure 5.13: The overview for case study 1. In this case study, we focus on finding unusual rises in glucose readings.

enable the user to quickly and easily find outlier events and their causes. We quickly find and select one child in the Time Series Map because it is the only instance that contains readings in a dangerously high range. See Figure 5.13 and the accompanying video. We use the day-oriented view to identify the specific hours of these spikes for further exploration. Combined with the detailed event logs [251], the three occurrences can be interpreted as follows:

- On 7 Sep at 12:00: The reading shows a sharp rise immediately after the patient's lunch, which included carrots, tofu, tortillas, and cream cheese.
- On 8 Sep at 11:50: The reading rises again, presumably for the same reason, as lunch was 10 minutes earlier than the previous day.
- On 18 Sep at 20:20: The event log does not show any recorded activity. However, the glucose monitor appears to malfunction, as the readings consistently show values of 396.396 and 401.

In this case, we effectively use Time Series Maps to identify an unusual sharp rise as it is the only instance of dangerously high readings. The day-oriented view also reveals unusual sharp falls, which are caused by the disconnection of the CGM.

**Case Study 2: Observing Daily Patterns:** In this case study, we observe the repeated appearance of patterns in glucose readings. Identifying patterns such as consistent increases or decreases in blood glucose levels can inform adjustments in medication, diet, or lifestyle to achieve better overall control. On many days, a prolonged period of high glucose levels is observed, which lasts approximately 8 hours between 3am and 11am. This is known as the dawn phenomenon, which results from natural hormonal changes. During these times, the patient’s glucose readings consistently fall within high and very high ranges, indicating that insulin dosage is insufficient. This is especially evident during the four days, when the readings remain consistently in the very high range, as shown in [Figure 5.12](#).

**Case Study 3: Finding Hypoglycemia:** In this case study, we use Time Series Maps to identify hypoglycemia. Our analysis incorporates glucose readings alongside event log and insulin records from the data set [251]. Hypoglycemia occurs when the blood glucose level is below 70 mg/dl. This condition needs immediate treatment. In [Figure 5.14](#), we begin by investigating the child nodes with a higher concentration of colours representing low and very low ranges. As there are unusually low readings from the data set (consistent readings of 5), we believe these readings are erroneous and exclude them. We then utilise the details-on-demand view to identify reverse bell curves where the readings consistently fall within the low or very low range, leading to a match on 17 September between 12:30pm and 1pm. There is no recorded food intake in the event log provided for that period. Upon reviewing the insulin records, a dose of bolus (quick-acting) insulin was administered at 10:33 am, approximately 82 minutes earlier than the usual time, according to the insulin records.

### 5.5.3 Health Data Experts Interviews

We conducted semi-structured interviews with each health data expert group and explained and demonstrated each aspect of our approach to obtain qualitative feedback and expert evaluations. Feedback and evaluation sessions were organised as follows:

- We asked the experts to describe their backgrounds and why they originally started studying health data
- We asked them how they currently obtained an overview of time series data

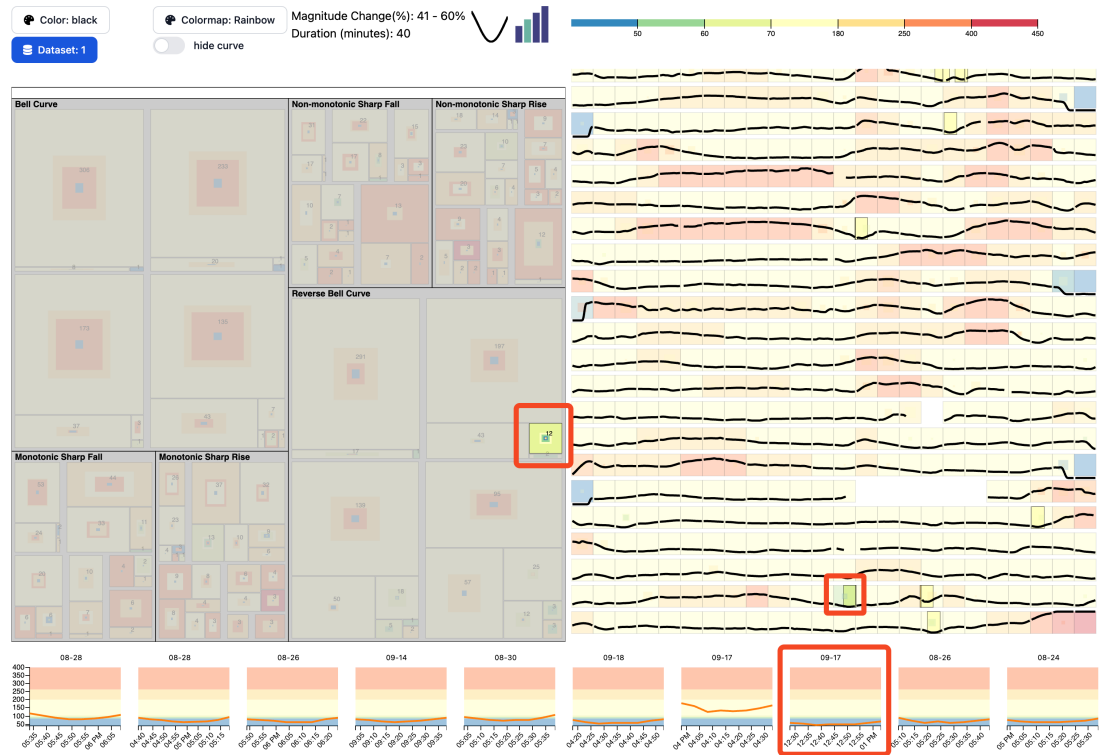


Figure 5.14: The overview for case study 3. In this case study, we focus on finding hypoglycemia, where glucose readings fall below 70 mg/dL.

- We asked if they are familiar with treemaps
- We explained the idea behind our approach starting with the EoIs, 2D matrices, and event hierarchy
- We demonstrated the Time Series Map view soliciting feedback and discussion
- This was followed by a demonstration of the day-oriented and detailed views soliciting feedback and discussion

A total of four health data experts were involved in this evaluation process, which was conducted individually for each expert group to ensure relevant and focused feedback. The interviews were conducted separately in two groups.

## Expert Group 1

**Background:** Group 1 consists of two health data experts. Expert 1 (E1) is an Assistant Professor at Northeastern University, with over 10 years of experience in visualisation of time series health data, with a focus on diagnostic and treatment decision support systems for diabetes and neurological conditions. The expert's child was diagnosed with type 1 diabetes in 2016, and the lack of competent tools to analyse diabetes

data inspired their research interest in blood glucose management. E1 focuses primarily on type 1 diabetes research. E1 developed their interest in health care with a funded PhD research assistantship, working on visualising networks of patients and concepts related to their care, with data extracted from patient discharge summaries. They then worked at IBM Watson Health which a focus on exploring the use of visualisation for health data.

Expert 2 (E2) is a postdoctoral researcher at the University of Konstanz, with expertise in health visualisation, specifically in the analysis of blood glucose data. E2 started studying blood-glucose related data in 2018, as a practical application of temporal event sequence visualisation research. They previously worked on how graph layout algorithms can be used to create overviews of Type 1 Diabetes data.

**Feedback:** When demonstrating the Time Series Map view and its interactions with other views, E1 provided the following feedback:

*“Overall, I really like this idea of separating out the patterns and clicking on patterns to see everywhere it occurs in all of the days, and getting a sense of the distribution of the different ranges within each of these patterns. I can imagine a clinician looking at sharp falls or just falls in general, and then drilling down into the ones that result in extreme lows.”*

E2 adds: *“I think reconciling overview visualisation with detail visualisation is an extremely complicated, still unsolved problem. Highlighting patterns can absolutely help on this, to avoid having the attention of the reader wandering through uninteresting details and wasting precious time and attention. In a medical setting, this can be particularly important, especially considering the short time doctors have to evaluate large amounts of data.”*

In addition, E1 commented on the Time Series Map view:

*“I do see value in seeing extreme readings. The green samples are not interesting, these are not something that you would treat. But the bell curve that goes up and down is definitely something I’d want to pay attention to. The bell curve can be used more directly in the case of meals or other events. The reverse bell curve, I am not sure if it’s long enough duration (150 mins), is useful for seeing the period of exercise.”*

When asked about potential use cases for the day view, E1 gave the following



feedback:

*“I think there is value in searching for patterns, and this is so different from what the clinicians are used to. The clinicians usually look at 2 weeks’ worth of data. The real power of this comes when you are trying to do a search across many days. Yes I do see the possibility of this being valuable.”*

E2 adds: *“I think this could let a clinician be able to read and have a general understanding about more data at a time—more than 2 weeks of data. In our previous research, we did try to focus on visualising patterns over multiple days for exactly this purpose, however, the approach we used was completely different. We did notice, though, that highlighting patterns can help spot systematic issues in the behaviour of patients.”*

**Expert Guided Features:** When asked about changes that can be made to improve our implementation, E1 responded:

*“A percentile chart to show the distribution of the readings could be useful, especially for observing variability.*

*If you’re going to keep that you know primary goal of identifying common and uncommon events, like with the area encoding here, I’d suggest removing the lines from the left from the treemap, because due to the area encoded in the treemap, it is difficult to see the lines within the nodes.”*

Based on E1’s suggestion on distinguishing extreme glucose levels, we added another colormap to help clinicians make more informed and rapid decisions to ensure patient safety. We also added a user option to turn representative curves on and off. See [Figure 5.15](#).

*“I think there would be some good value in having a diverging colour scale, or something where it’s clear that the extremums are different, because the way that a diabetes clinician will treat those extremums is very different. For example, for most patients, the clinician would strive to virtually eliminate lows, especially if there’s any concern that the patient or their caregivers can’t address it rapidly enough to keep the patient out of the hospital, or keep the patient out of a coma, or from needing glucagon to bring them back up. So they’re much more tolerant of high excursions.”*

When looking at the interaction between the Time Series Map view and the day

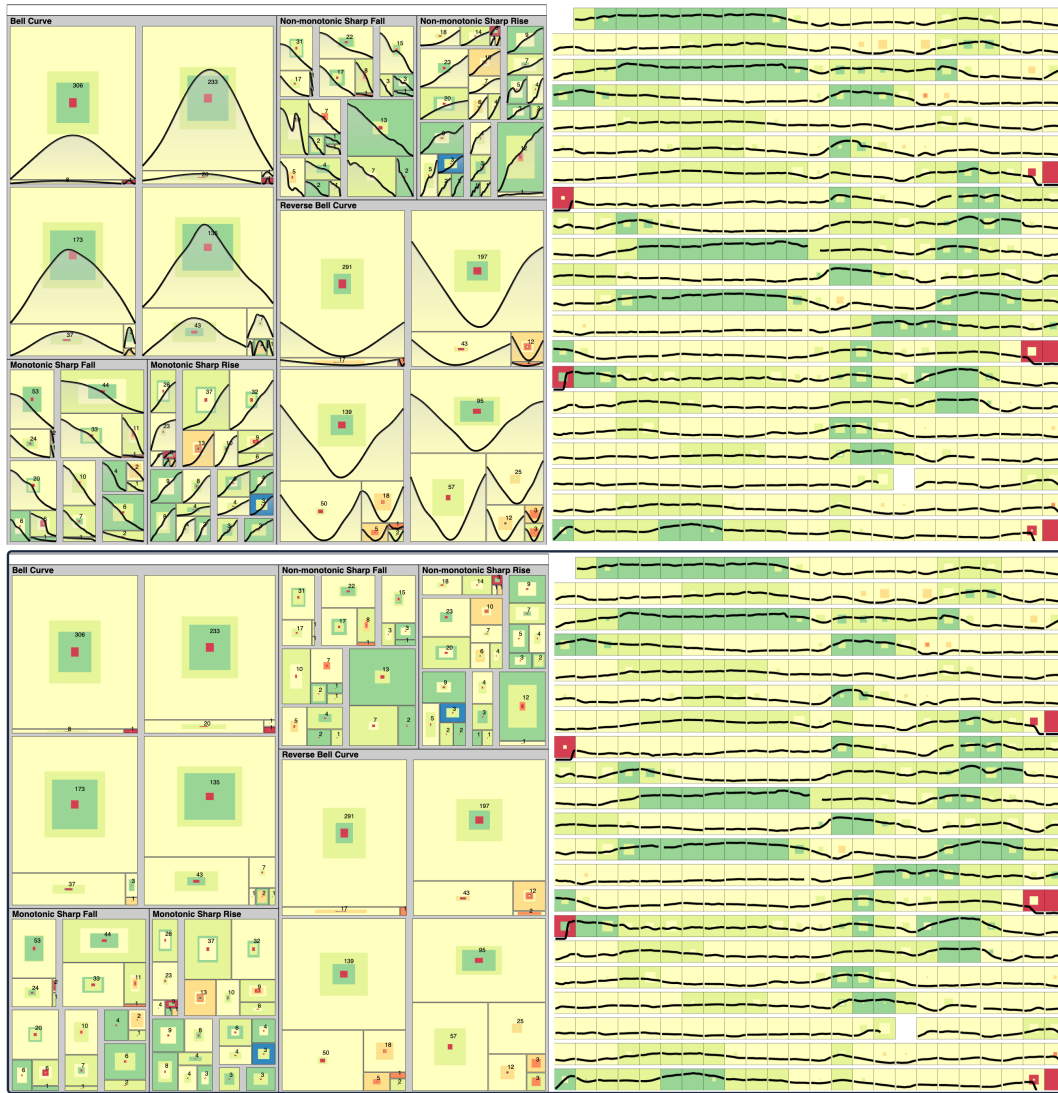


Figure 5.15: Comparison of the Time Series Map view, with representative curves enabled (top) and disabled (bottom), using the colormap suggested by health data experts, which is derived from ColorBrewer [28].

view, E1 made the following suggestion:

*“From a health informatics standpoint, where maybe you’re looking at multiple patients and you’re trying to understand how they behave on a new drug you’re trying, or whether there’s general trends across patients, that could be an interesting application. I’d love to see an example of this with fewer days just to get a better sense of what the treemap would look like.”*

E1 commented that the representative curve in the Time Series Map view is a good idea:

*“These curves they are useful, but they block the nodes behind them and make the*

*node-to-node comparison difficult.”*

In response to this feedback, we introduced an option to toggle the visibility of representative curves in the Time Series Map view.

E1 also thought that our implementation is novel therefore a comprehensive onboarding process is necessary:

*“I have a feeling, for clinicians, it would have to be a long study with lots of training because this is so different from what they’re used to. They’re not used to searching for patterns, they’re used to doing a manual visual search across 14 days.”*

E2 adds: *“The applications of this methods could extend to more than this domain. I am thinking about how I track my daily productivity trying to find low productivity days so I can avoid the events that cause them.”*

## **Expert Group 2**

**Background:** Group 2 consists of two health data experts. Expert 3 (E3) is a Research Fellow at University of Birmingham, specialising in patient self-management, health service improvement, and digital health, with a focus on communications via readily accessible and understandable visualisations. E3 has over twenty years of experience in researching health data collected routinely and prospectively in primary and secondary care settings, with a focus on improving the effectiveness of health service delivery. E3 developed their interest in blood glucose data as part of an ongoing research programme funded by the UK’s National Institute for Health and Care Research, with a focus on children and young people’s blood glucose monitoring and control.

Expert 4 (E4) is an assistant professor at Aston University, specialising in clinical care processes and visualisation. E4 has been researching data mining and analytics since 2016, with experience in applying process mining techniques to understand time-based patterns of drug prescription and their alignment with multimorbidity. E4 developed their interest in combining visualisation of blood glucose data with process mining algorithms to inspire more informed analysis of clinical processes.

**Feedback:** When viewing the Time Series Map view, E3 immediately made the following observation:

*“From the visualisation, I can immediately see the amount of green, where it’s*

*actually staying within range. That is important. I think it's quite powerful, the fact that you can break the whole data set down and at a glance, as a clinician, you can see the practical patterns and you want to pass on to the patient.*

*I actually think it could be as equally powerful if it was patient-facing. Of course they need to understand exactly the type of curve and whatever per se. But in terms of an overview, the patient could see there's that issue on a Thursday and you could do something about it, because the issue might not stand out if you are just looking at the readings alone."*

After our complete demonstration of all views and features, E3 commented on our implementation:

*"The continuous glucose monitor generates a lot of data, this overview offers the benefits of understanding the data at a glance. In healthcare, for shared decision-making to happen the patient needs to be empowered by information to understand and make informed decisions, and I can see this interface could be useful for that."*

E4 provided the following feedback when seeing the interactions, specifically between the Time Series Map view and the day view, and between the Time Series Map view and the detail view:

*"The treemap provides a good way of seeing the overview of the patterns within the data, but we don't know about how all those patterns occur in time. Then on the right-hand side you can start drilling into particular things. So if we want the details of the curves we would click on one of the rectangles in the treemap and then get the actual details and this is where we would find it useful."*

When asked about the potential use cases for our implementation, E4 responded:

*"I can see it being useful, compressing arbitrary length sequences down and extracting the main features from that. An end user could potentially benefit from seeing colours and shapes, they give a very intuitive early warning sign potentially."*

**Expert Guided Features:** E3 suggested a filtering feature and a bidirectional interaction between the Time Series Map view and the day view:

*"Filtering by time of day would be useful, as it would allow us to see patterns at different times of the day, such as between 4 am and 6 am. Bidirectional interaction would be useful, as it would allow us to click on the day view to see the patterns in the*

*treemap.* ”

Similarly to E1, E4 suggested that we need to carefully design the onboarding process to ensure first-time users understand how to navigate through all the features provided by our implementation.

*“The interface might seem visually too busy at first, as it condenses a massive amount of information into something that’s small and fixed dimension. You need to walk the user through it, adapt the interface per use case, before they find it useful.”*

## 5.6 Limitations and Future Work

The current implementation has been tested on data sets with 68,000 events over eight months, we further tested with a synthetic data set three times larger and did not observe any performance issues. However, scalability with extremely large data sets, such as millions of events, likely needs to be managed with desktop applications instead of a web application, or with a specialised client-server implementation.

To increase the applicability of the hierarchical visualisation approach, future research should explore its adaptation to other domains with different types of time series data. This includes identifying relevant event categories and designing appropriate visual representations for these domains.

Our approach is novel and differs significantly from the traditional time series visualisation techniques to which clinicians and patients are accustomed to. User training is required to ensure effective use and help users understand the potential of our design in their workflows.

## 5.7 Conclusions

In this paper, we introduce Time Series Maps, a novel hierarchical visualisation approach for managing and exploring long time series data with a focus on continuous blood glucose readings. Time Series Maps integrate the strengths of time series and event sequence visualisation to provide a comprehensive, scalable overview of blood glucose data, aiding in the identification of significant patterns and events.

By structuring long-term patient histories into visually navigable forms, this work

contributes to the broader “EHR Vis” framework by addressing the challenge of temporal abstraction, complementing [Chapter 3](#)’s transformation of unstructured text and [Chapter 4](#)’s incorporation of geospatial context. Just as clinical narratives require structured representation and spatial data must maintain geographic coherence, temporal visualisations must preserve the integrity of event sequences while enabling scalable exploration.

We demonstrated the effectiveness of Time Series Maps on two real-world continuous glucose monitoring data sets, showcasing its ability to generate overviews for extended data sets and support user exploration through interactive visual designs. The hierarchical structure allows users to explore data at multiple levels, from high-level overviews to detailed examinations of individual events. This flexibility is particularly beneficial for clinicians and researchers needing to identify trends, anomalies, and key events within long time series data sets.

This chapter contributes to the broader EHR Vis narrative by transforming vast, multidimensional data into accessible visualisations without losing the depth required for clinical insight.

## Chapter 6

# EnsembleDashVis Views and Volunteers - A Retrospective and Early History

Wang, Q., Borgo, R., & Laramée, R. S. (2024). EnsembleDashVis Views and Volunteers – A Retrospective and Early History. In M. Bassanello, R. Geppini, X.-N. Li, & A. Matecki (Eds.), *New Community Health Models*. IntechOpen. <https://doi.org/10.5772/intechopen.115029> [318]

*“Above all else, show the data.”*

– Edward R. Tufte, the Father of Information Design (1942 - present)

The chapter is based on the book chapter published in *New Community Health Models* [318].

This chapter is an unconventional chapter and represents a unique, unplanned experience during this Ph.D. The chapter offers a retrospective history of the early development stages of EnsembleDashVis, a visualisation dashboard specifically crafted to support modellers in interpreting a simulation model utilised to forecast COVID-19 trends. The volunteer effort behind this dashboard was collaboratively contributed with the Scottish COVID-19 Response Consortium (SCRC), with the objective of enabling an enhanced comprehension of the complex dynamics of the pandemic through the modelling of COVID-19 data collected by the National Health Service (NHS) during the first wave of the outbreak in Scotland.

This retrospective chronicles the design and development journey of the system, guided by feedback from domain experts, all taking place amidst the exceptional circumstances of an unprecedented pandemic. Our expertise in EHR Vis was able to contribute significantly to combating the pandemic, and the lessons learnt from this experience have been invaluable in shaping the future of cross-disciplinary collaborations in the field of Visualisation and beyond.

## Contents

---

<b>6.1</b>	<b>Introduction and Motivation . . . . .</b>	<b>175</b>
<b>6.2</b>	<b>Background and Related Work . . . . .</b>	<b>177</b>
6.2.1	VIS for Emergency Response . . . . .	178
6.2.2	VIS for COVID-19 Data Modelling . . . . .	178
<b>6.3</b>	<b>Data Description . . . . .</b>	<b>180</b>
<b>6.4</b>	<b>EnsembleDashVis . . . . .</b>	<b>182</b>
6.4.1	An Unconventional Software Development Cycle . . .	183
6.4.2	Technology and Design . . . . .	184
6.4.3	Interaction . . . . .	185
6.4.4	Meetings and Milestones . . . . .	186
<b>6.5</b>	<b>Domain Expert Feedback . . . . .</b>	<b>194</b>
6.5.1	Summary of Feedback . . . . .	194
6.5.2	Detailed Feedback . . . . .	195
<b>6.6</b>	<b>Limitations . . . . .</b>	<b>197</b>
<b>6.7</b>	<b>Conclusions . . . . .</b>	<b>199</b>

---



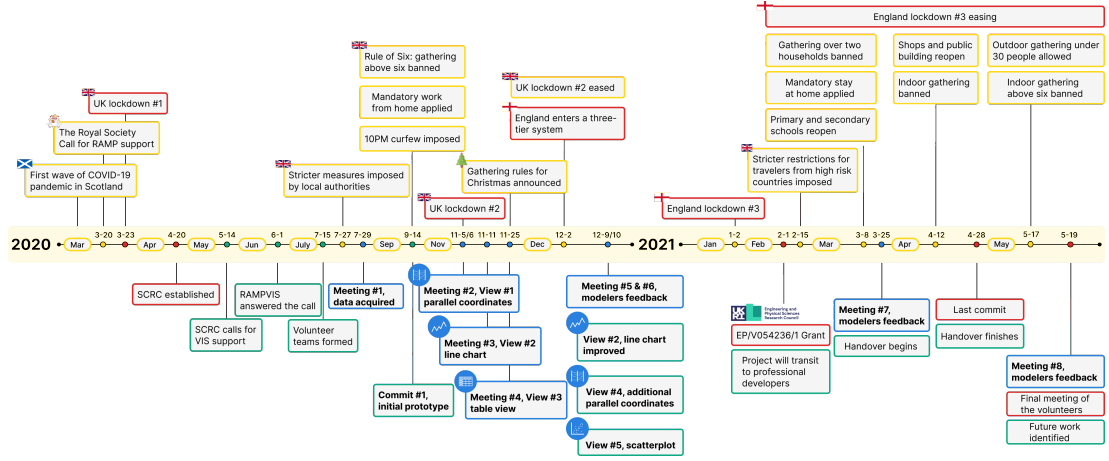


Figure 6.1: A timeline of the events between Mar 2020 and the end of our volunteer work on 19 May 2021. The upper section include **policy changes** during the time span, the lower section includes **project developments** and **meetings**. **Milestone** events are shown in red.

## 6.1 Introduction and Motivation

The Scottish COVID-19 Response Consortium (SCRC) [279], in collaboration with the Royal Society’s call to action in March 2020, has taken a proactive approach to address the need for enhanced epidemiological models of COVID-19 transmission. This joint volunteer effort, known as Rapid Assistance in Modelling the Pandemic (RAMP) [271], aims to foster a deeper understanding of the consequences associated with various exit strategies from lockdown measures. Moreover, this consortium attracted the involvement of distinguished scientists and experts from diverse organisations both within the United Kingdom and abroad, thus augmenting the collective knowledge base and ensuring comprehensive expertise in specialised domains.

RAMPVis [281] is a group of researchers specialised in Data Visualisation and Visual Analytics (abbreviated as VIS). The group voluntarily came forward to contribute its specialised skills and knowledge in order to provide valuable support to the SCRC modellers. The term *modellers* used here refers to the SCRC researchers who were actively engaged in the development of epidemiological models in SCRC.

This target user group predominantly includes experts in domains such as mathematics, statistics, and epidemiology.

Serving as the volunteer team responsible for providing visualisation support to one of the epidemiological models developed by the SCRC modellers [301], our main

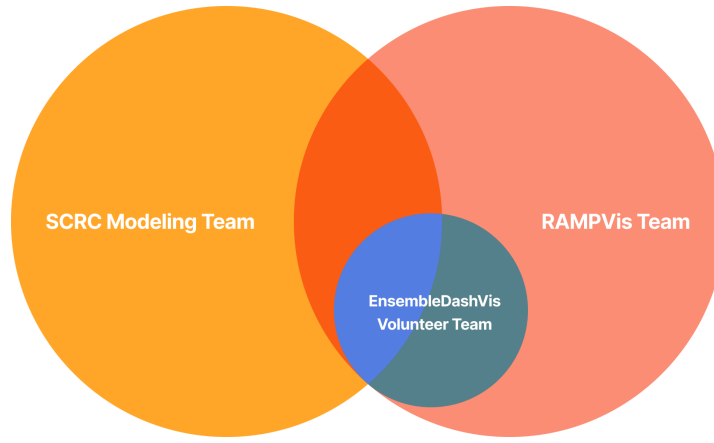


Figure 6.2: The organisation of researchers from the SCRC and RAMPVis. The SCRC modelling team is responsible for developing the epidemiological models leveraging different modelling techniques. The RAMPVis team provides visualisation support to the SCRC modelling team, by establishing four VIS volunteer teams who work on the actual development under the guidance of the RAMPVis team.

objective is to provide Data Visualisation and Visual Analytics (VIS) researchers and practitioners with valuable insights gained from our research and development (R&D) activities conducted during the COVID-19 pandemic. In an effort to predict the potential impact of diverse interventions, modellers have actively utilised COVID-19 data, employing a method known as Uncertainty Quantification (UQ). This process seeks to measure uncertainties through the application of mathematical models and simulations. However, modellers are faced with significant challenges, including the aspects of expert elicitation and effective communication. In other words, there is a need for software engineering efforts coupled with visualisation to provide support for validation and verification tests of models, and to create efficient workflows between modellers and researchers from other disciplines [299].

In addressing these hurdles, VIS emerge as a potent tool, offering the capacity to significantly enhance and streamline their collaborative workflows [310]. While our work may not have showcased the state-of-the-art VIS techniques, it effectively delivered rapid and practical VIS support to the modellers during an exceptional and demanding time.

Our contribution is an early history of our volunteer response from a software engineering and visualisation perspective. We present the earliest stages of the visualisation dashboard, EnsembleDashVis, developed during the pandemic, aiming to assist the modellers in interpreting an Approximate Bayesian Computation Sequential Monte

Carlo (ABC-SMC) inference model that they have developed using COVID-19 data collected during the first wave of the outbreak in Scotland [275]. Much of this effort and the reasoning behind this volunteer work was never documented.

**Unconventional Software Development:** The visualisation software created in this project was developed under unconventional and unprecedented circumstances.

One of the distinctive features of this software project was the significant level of uncertainty encountered at the project’s inception. The following aspects were unknown at the project outset:

- An unknown a priori requirements specification: We did not know what the user requirements and expectations were.
- An unknown project team: The members of the project team were unknown and/or had no previous history of collaboration. We only knew the leader of the visualisation team, Prof. Min Chen. In addition, the project team was dynamic, with new members joining throughout.
- Unknown data characteristics: We did not know what the simulation data was at the start of the project.
- An unfamiliar work environment: The landscape of the collective work environment changed to a work-at-home model, which was new to the team at the time.

While arguably, these characteristics could describe other software engineering projects, we believe that the uncertainty in this particular case was unusually high. All aspects of this project had the feel of *“laying down the tracks as the train was running”*.

## 6.2 Background and Related Work

VIS has been widely utilised in critical applications such as emergency responses and healthcare, assisting public officials and decision makers in understanding intricate data sets and extracting useful, actionable insights from them [160]. VIS has also played a prominent role in disseminating COVID-19 information through various media channels. It has played a substantial role in enhancing public communication, making it more efficient and clear, thereby fostering a wider comprehension of the crisis [346].

In our work, our primary objective was to extend support through VIS to two

distinct user groups. Firstly, the statisticians, who could significantly benefit from VIS in comprehending their models more effectively and fine-tuning them. Secondly, to the epidemiologists, whom VIS could assist in interpreting the outcomes of these computational models. Our outcomes are later included in multiple publications [301, 302, 305, 306]. The early stages functioned as the preliminary VIS prototype, shaping a portion of their respective studies. We refer the reader to Preim and Kai for an overview of VIS applied in the context of public health [270].

### 6.2.1 VIS for Emergency Response

Previously, we described related work that focuses on the use of VIS in emergency response. We refer readers to the related work section in Chen et al. [301]. The aforementioned literature review laid the foundation and was conducted prior to the development of our study in 2020.

Maciejewski et al. [97] develop a VIS toolkit to analyse the effect of decision measures enforced during a simulated pandemic, the tool was later utilised by the Indiana State Department of Health during an outbreak of H1N1 (swine flu). Ribicic et al. [112] leverage VIS with the intention of delivering real-time feedback derived from flood simulations to nonexpert users, while Konev et al. [135] use VIS to support decision-making in flooding scenarios.

Jeitler et al. [233] use VIS to analyse social media data to aid rescue teams, specifically in terms of optimal allocation of resources during emergency response situations. Similarly, Nguyen and Dang [242] harness social media data, paired with VIS, to facilitate and improve post-earthquake resource allocation and rescue effort.

In contrast to the majority of previous studies mentioned here that generally focus on preparing for future emergencies, our work was undertaken during the COVID-19 pandemic as a rapid response to a then current and ongoing emergency.

### 6.2.2 VIS for COVID-19 Data Modelling

In the rest of the section, we focus on the use of VIS to analyse the computational modelling of COVID-19 data. These studies were not published nor available to us during the development of the work we present here (from July 2020 to April 2021).

In fact, the use of VIS in epidemiological modelling was rare, the modellers might have not known that they had such a potent instrument readily available [301].

He et al. [262] developed an SEIR (Susceptible, Exposed, Infected, and Recovered) model for spread prediction by leveraging COVID-19 data obtained from the Hubei province in China. They employed a variety of 2D plots to estimate the parameters of the model and interpret the results that the model yielded. Godio et al. [260] took the same approach in developing an SEIR model for the Lombardy region in Italy.

The IHME COVID-19 Forecasting Team [ihmecovid-19forecastingteam2021Modelling] take the application of data visualisation (VIS) a step further in their development of the SEIR model for accessing social distance mandates, they extend the use of VIS to include choropleth and violin plots, and small multiples for 2D plots.

Chinazzi et al. [255] develop a model to simulate the effectiveness of international travel restrictions in containing the spread of COVID-19. In addition to the use of 2D plots to refine their models, they also utilise a range of geospatial approaches. This enabled them to more effectively interpret the results generated by their models. The use of geospatial visualisations is also adopted by Alvarez Castro and Ford [285] in their development of a model for analysing transmission in a university campus in the UK.

Studies have also been introduced which focus on the individual level, examining the transmission chain from person to person. Antweiler et al. [286] collaborated with public health departments in Germany and introduced a novel visual analytic method to identify clusters of COVID-19 infections in contact tracing networks. Meanwhile, Baumgart et al. [288] presented a visualisation system designed to explore and analyse the pathways of pathogen transmission within hospitals. The system leverages linked views, including a transmission pathway view inspired by storyline visualisation, aiming for efficient and intuitive contact tracing.

In contrast to these studies that highlight the efficacy of VIS in supporting the computational modelling of COVID-19 data with a primary focus on model development, as they are formulated by the modellers, our study takes a different approach. We focus our attention on exploring VIS as a potent tool that can significantly improve the computational modelling of COVID-19 data, all viewed through the unique lens of a VIS practitioner.

Table 6.1: 16 input parameters for the ABC-SMC inference model. As constant parameters such as  $K$  and  $rrd$  do not affect the simulation results, they are not rendered in our visual designs.

Name	Description
T_lat	Mean latent period (days)
juvp_s	Probability of juvenile developing symptoms
T_inf	Mean asymptomatic period (days)
T_rec	Mean time to recovery if symptomatic (days)
T_sym	Mean symptomatic period prior to hospitalisation (days)
T_hos	Mean hospitalisation stay (days)
inf_asym	Reduction factor of infectiousness for asymptomatic infectious individuals
p_inf	Probability of Infection
p_hcw	Probability of Infection (Healthcare Worker)
c_hcw	Mean number of Healthcare Worker contacts per day
d	Proportion of population observing social distancing
q	Proportion of normal contact made by people self-isolating
p_s	Age-dependent probability of developing symptoms
rrd	Risk of death if not hospitalised
lambda	Background transmission rate
K	Hospital bed capacity

## 6.3 Data Description

The data used in our work includes simulation parameters and outcomes from an ABC-SMC inference model [61] developed by a group of modellers from Durham University, the University of Edinburgh, the University of Exeter, the University of Glasgow, and the London School of Hygiene & Tropical Medicine. The pandemic data used for the simulation was collected by NHS Scotland during the first wave of the outbreak in Scotland spanning a period of 59 days [275].

The model was built to analyse the pandemic data and infer the parameters of the model that best fit the data. The model accepts 16 input parameters (see Table 6.1), and a random seed facilitates the generation of 160 distinct sets of configurations for these input parameters. The model then employs these configurations as the initial input to perform 1,000 simulation iterations. As the outcome of these simulations, 160 sets of predictions are generated, each containing 13 output parameters, as shown in Table 6.2.

Upon receiving the data, we consulted the modellers to gain insights into the con-

Table 6.2: 13 output parameters from the simulation performed by the ABC-SMC inference model.

Name	Description
iter	The simulation number.
day	The day number.
age_group	The age group of the population.
S	Number of susceptible individuals (not infected).
E	Number of infected individuals but not yet infectious (exposed).
E_t	Number of exposed individuals and tested positive.
I_p	Number of infected and infectious symptomatic individuals but at pre-clinical stage (show yet no symptoms).
I_t	Number of tested positive individuals that are infectious.
I	Number of infected and infectious asymptomatic individuals.
I_s	Number of infected and infectious symptomatic individuals.
H	Number of infected individuals that are hospitalised.
R	Number of infected individuals that have recovered from the infection.
D	Number of deceased individuals due to the disease.

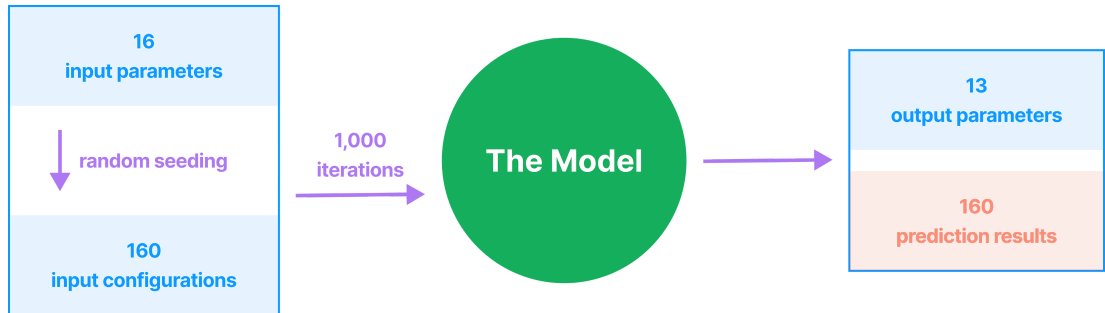


Figure 6.3: An illustration of the flow from the input parameters to the prediction results. 160 sets of input parameters are used to perform 1,000 simulation iterations, resulting in 160 sets of prediction results.

ventional workflow they employ for data processing, as well as the significance and the underlying meaning associated with each input and output parameter. As constant parameters such as  $K$  and  $rrd$  do not affect the simulation results, they are not rendered in our visual designs.

It is worth mentioning that after plotting the output data using a line chart, an error was immediately spotted, see [Figure 6.7](#), where an unusual spike can be observed on day 20. The modellers were notified and the bug was fixed. However, the rectified output file was never made available to us.

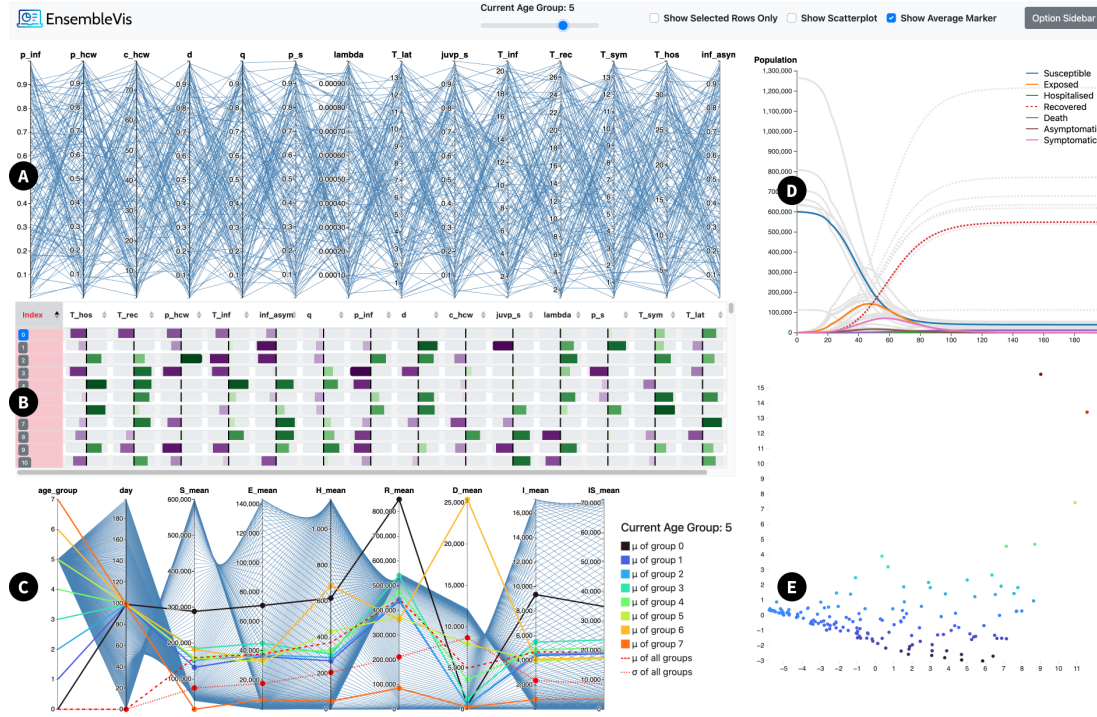


Figure 6.4: The overview of EnsembleDashVis. The dashboard consists of five views: (Figure 6.4A) a parallel coordinates plot for all input configurations, (Figure 6.4B) a table view with glyphs for all input configurations, (Figure 6.4C) a parallel coordinates plot with brushing to enable quick simulation outcomes filtering, (Figure 6.4D) a line chart for model predictions, and (Figure 6.4E) a scatterplot for Principal Component Analysis (PCA) outcomes. The views are coordinated with each other, enabling the modellers to observe relationships between input and outcome through interactions.

## 6.4 EnsembleDashVis

This section presents the development of EnsembleDashVis from its technology and design and interaction techniques. We then present the history behind our fully virtual collaboration between volunteer researchers from multiple UK institutions. Being one of the four VIS volunteer teams, we received guidance from the RAMPVis team through regular virtual meetings. The RAMPVis team regularly communicated with the SCRC modelling team and provided us with important information and data. We chronicle the development of different views of the data, the order in which they were introduced, and the reasons and motivations at the time. In 2020 we were all in an unprecedented and unfamiliar situation, thus, some of our decisions were ad hoc.



### 6.4.1 An Unconventional Software Development Cycle

A common agile software development life-cycle consists of five stages: 1) requirements specification, 2) software design, 3) implementation, 4) testing, 5) documentation. [5] And these five stages iterate repeatedly until the software project is finished. However, this project deviated significantly from the standard agile software engineering model.

**Knowledge Exchange:** This project, as well as all other visualisation projects we have collaborated on, starts with a phase more appropriately named *Knowledge Exchange (KE)*. This is due to the fact that the domain experts do not have a background in visualisation, thus they do not know what the options are in terms of visual analysis. As a result of this absence of visualisation expertise, the KE phase (which replaces the standard requirements specification phase) involves two sub-phases:

**From Domain Experts to Visualisation Team:** The discussion starts with the visualisation team asking a series of questions to the domain experts. These questions are typically:

1. What data have you collected?
2. Why did you collect the data?
3. What questions were you trying to answer with the (simulation in this case) data?
4. What information were you hoping to obtain as an outcome from your data collection process?
5. Can you describe the characteristics of your data in more detail? After the visualisation team has gathered enough of the first round of knowledge, the next phase of the KE process can begin.

**From Visualisation Team to Domain Experts:** Since the domain experts, in this case the simulation experts, do not have a background in visualisation, they look to the visualisation team to *make recommendations* to them in terms of what visual analysis designs might make sense. Thus, the visualisation team typically discusses options in terms of graphical displays that might help the domain experts answer the questions posed in the previous sub-phase. In essence, the KE process flows in the other direction. The visualisation team essentially educates the domain experts on visual analysis options that they may not be familiar with. After this discussion, the

actual next phase of the software engineering lifecycle can begin.

The software design and implementation phases are the same as in the typical agile model of software development.

**Testing and Evaluation:** Instead of the conventional testing phase of a typical agile development model, this project and our other collaborative visualisation projects, undergo a more appropriately described *testing and evaluation (TE)* phase. Instead of the emphasis on extensive testing on a wide range of cases, our visualisation software undergoes an extensive evaluation by the domain experts. Specifically, they carefully evaluate if and how the software can be used to answer their domain-specific questions or hypotheses. They will ask for a demonstration of precisely how it can be used for their specific application. Typically, when we demonstrate a version of the visualisation software, the domain experts will ask several questions about how it works. And then, during the discussion new feature requests arise. Often these sessions are also characterised by feature creep [50]. The TE phase is usually fairly intense generating a lot of enthusiasm from the domain experts since they are seeing visualisation software that they have never seen before and thus a large number of feature requests arise from the meetings in this phase.

After the TE phase the cycle repeats interactively. In the visualisation software development lifecycle, the requirements specification phase is replaced by the KE phase and the testing phase is replaced by the TE phase. This is because adequate knowledge transfer and evaluation cannot be completed in one single cycle. The cycle repeats until the project ends, typically constrained by a funding period.

### 6.4.2 Technology and Design

The development of EnsembleDashVis was carried out using a combination of web technologies, including HTML, CSS, and JavaScript. The dashboard was designed to be a web-based application, enabling it to be accessed from any device with a web browser. The dashboard was built using D3.js [328], which is a powerful and flexible library for creating visual data representations in web applications. D3.js provides a wide range of tools for creating interactive graphics, including support for a wide range of data formats, and a large number of built-in visual designs. The dashboard was

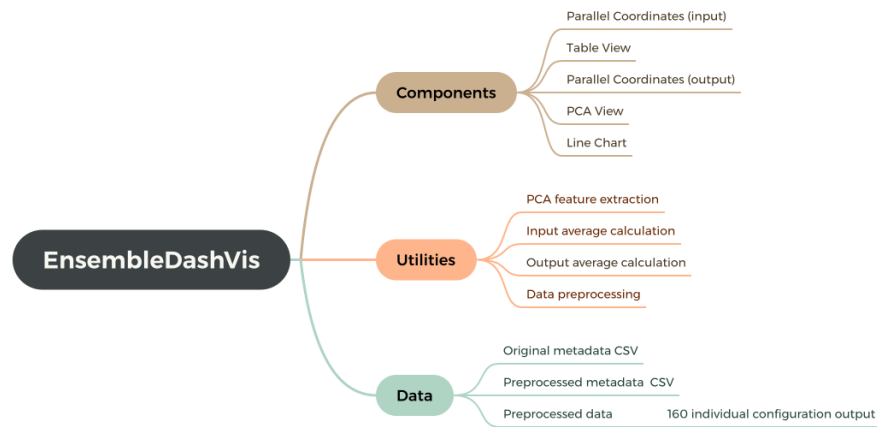


Figure 6.5: The structure of the actual code. Components are organised into separate files, with each file containing the code for a single view. Utilities contain the code for the data preprocessing and calculations. Data contains the metadata and preprocessed output by utilities.

designed to be responsive, allowing it to adapt to different screen sizes and orientations, and to be accessible, allowing it to be used by people with disabilities.

The dashboard was then hosted on Netlify [267], which provided unlimited credits to websites that were dedicated to sharing information about COVID-19. This allowed the dashboard to be accessed by anyone with an internet connection, which was crucial during the pandemic for virtual collaboration.

The dashboard was designed to be easy to use, with a simple and intuitive interface that enables users to quickly and easily explore the data. It employs a modular design, with each view of the data rendered as a separate component, allowing the dashboard to be easily extended and modified. Data is preprocessed by utility functions and stored in separate CSV files, which is then loaded into the dashboard when accessed.

The source code is publicly available on GitHub, <https://github.com/thevisgroup/EnsembleVis> [297].

### 6.4.3 Interaction

In this section, we describe the interaction techniques that were incorporated into the dashboard to enable the modellers to explore the data and identify interesting patterns. Here we follow the Visual Information Seeking Mantra [9]: “overview first, zoom and filter, then details-on-demand”.

## Overview First

Figure 6.4 shows the overview of the dashboard. The dashboard consists of five views: Figure 6.4A a parallel coordinates plot [15, 105] for all input configurations, Figure 6.4B a table view with glyphs for all input configurations, Figure 6.4C a parallel coordinates plot for simulation outcomes, Figure 6.4D a line chart for model predictions, and Figure 6.4E a scatterplot for Principal Component Analysis (PCA) [320] outcomes.

Each view provides an overview of the corresponding data, supporting the modellers to quickly identify interesting patterns and outliers.

## Zoom and Filter

The parallel coordinates plot in Figure 6.4A and Figure 6.4C allows the modellers to select a subset of input parameters via brushing to focus on interesting configurations. The table view in Figure 6.4B enables the modellers to sort configurations by individual input parameters via sorting. The scatterplot in Figure 6.4E enables the modellers to reduce the dimensionality and identify key parameters via brushing.

These interactions enable the modellers to quickly adjust the focus of the views and drill down into the details.

## Details-on-Demand

These views in Figure 6.4 are coordinated with each other, e.g., brushing on the input parallel coordinates plot in Figure 6.4A highlights the corresponding input configurations in both the table view Figure 6.4B and scatterplot Figure 6.4E. Focusing on a specific row in the table view Figure 6.4B renders the corresponding output data in both the output parallel coordinates plot Figure 6.4C and line chart Figure 6.4D.

These coordinated interactions enable the modellers to quickly identify interesting configurations and observe relationships between input parameters and model outcomes.

### 6.4.4 Meetings and Milestones

In this section, we provide a detailed history of meetings and development milestones. Section 6.4.4 shows the list of meetings held throughout the entire volunteering period,

Date	Attendees	Milestones
27 July 2020	Dylan Rees, Elif Firat, Hui Fang, Min Chen, Qiru Wang, Rita Borgo, Robert Laramee, Tom Torsney-Weir	Volunteer team established.
6 Nov 2020	Cagatay Turkey, Hui Fang, Qiru Wang, Rita Borgo, Robert Laramee, Tom Torsney-Weir	First prototype.
6 Nov 2020	Ben Swallow, Hui Fang, Qiru Wang, Rita Borgo, Robert Laramee, Tom Torsney-Weir	First prototype feedback
11 Nov 2020	Cagatay Turkey, Elif Firat, Hui Fang, Rita Borgo, Robert Laramee, Qiru Wang, Tom Torsney-Weir	6GB of simulation data received. Second prototype.
25 Nov 2020	Cagatay Turkey, Elif Firat, Hui Fang, Robert Laramee, Qiru Wang	Third prototype.
9 Dec 2020	Cagatay Turkey, Hui Fang, Robert Laramee, Qiru Wang	All views implemented.
10 Dec 2020	Ben Swallow, Cagatay Turkey, Hossein Mohammadi, Hui Fang, Janine Illian, Michael Dunne, Peter Challenor, Qiru Wang, Richard Reeve, Robert Laramee, Thibaud Porphyre	Presentation to modellers.
25 Mar 2021	Cagatay Turkey, Elif Firat, Hui Fang, Rita Borgo, Robert Laramee, Qiru Wang	Further feedback from modellers.
19 May 2021	Ben Swallow, Cagatay Turkey, Hossein Mohammadi, Hui Fang, Janine Illian, Michael Dunne, Peter Challenor, Qiru Wang, Richard Reeve, Robert Laramee, Thibaud Porphyre	Final presentation to modellers.

Table 6.3: The table shows the list of meetings held throughout the entire volunteering period, detailing each meeting’s date, the attendees, and the milestones accomplished.

detailing each meeting’s date, the attendees, and the milestones accomplished.

### Meeting #1 - July 2020

On 27 July 2020, amid the UK’s first national lockdown and stricter measures imposed by local authorities, we convened the initial virtual meeting with VIS researchers from King’s College London, Loughborough University, Swansea University, University of Nottingham, University of Warwick, and University of Oxford.

During the meeting, we received an overview of the SCRC and the responsibilities of the visualisation volunteer team. Our assigned task was to create visual interfaces for the model, for the purpose of enabling the modellers to analyse the outcomes of the model.

Following the initial meeting, we engaged in email correspondence with the modellers to delve into the visualisation requirements. The modellers shared a comprehensive list of parameters and model outcomes, along with the corresponding outcome data [275].

### **Commit #1 - Sep 2020**

We proceeded to create an initial prototype of the visualisation, which was subsequently reviewed by the modellers. Incorporating their input, we refined the prototype during our weekly internal discussions. On 14 Sep 2020, England introduced the ‘rule of six’, which banned any gatherings above six. On the same day, we made our first commit to a GitHub repository (<https://github.com/thevisgroup/EnsembleVis>), signifying the commencement of our development. At the same time, we began preprocessing the data. A week after the initial commit, the UK witnessed the implementation of additional restrictions, such as mandatory work from home and a 10PM curfew.

### **Meeting #2, View #1 - Nov 2020**

On 5 Nov 2020, the first day of the second national lockdown in the UK, we completed the first view of the simulated input parameters, a parallel coordinates plot. See Figure 6.6. We chose to use a parallel coordinates plot as it is a common technique for visualising multivariate data, and is particularly useful to explore relationships and patterns across multiple input parameters. Each axis in the plot represents an input parameter, the y-axis represents the value of the parameter, and each polyline represents one input configuration. The plot supports brushing and linking, enabling modellers to select a subset of input parameters to focus on interesting configurations. This followed by the second meeting with the RAMPVis team from other institutions, where we received feedback on the first view, on 6 Nov 2020. The response from the modellers to the parallel coordinates view was, in general, very positive. They are very interested in multivariate analysis and had not seen this visual representation before. More details are provided in Section 6.5 on domain expert feedback.

### **Meeting #3, View #2 - Nov 2020**

On 11 Nov 2020, the group convened for the third meeting, where we received further feedback from the RAMPVis team on the parallel coordinates plot. As per

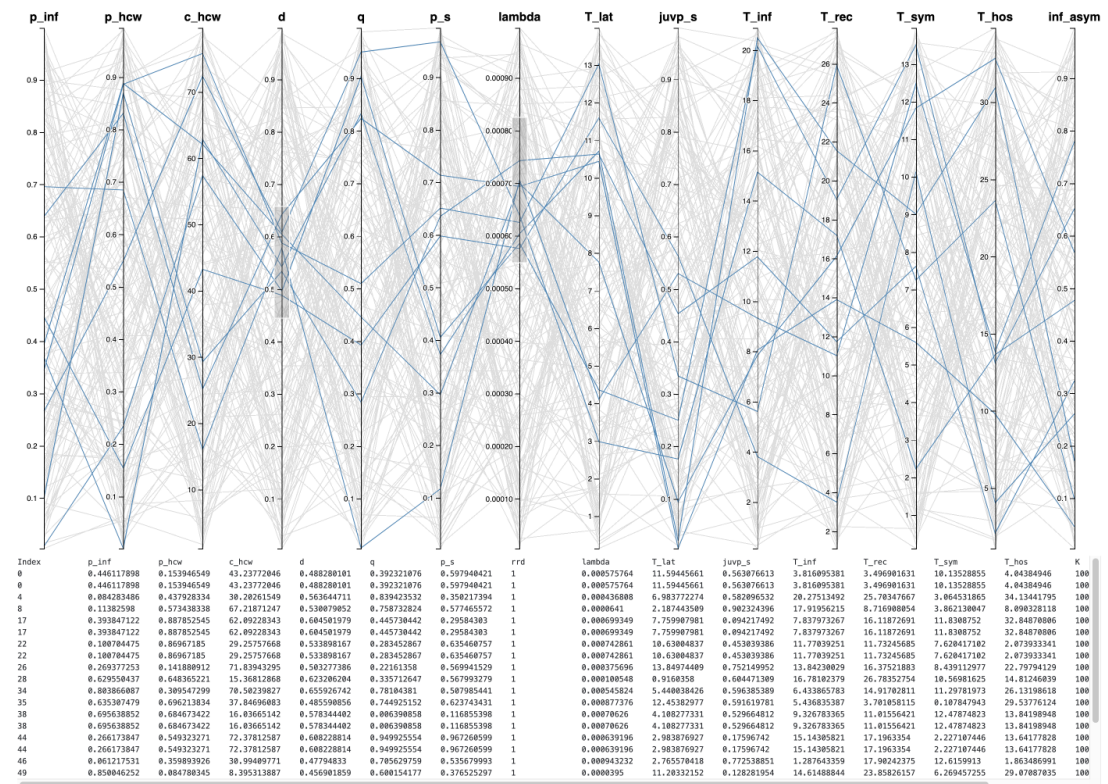


Figure 6.6: The first visual design, a parallel coordinates plot depicting all 160 input configurations of the model, was completed on 5 Nov 2020. Each axis represents an input parameter, the y-axis represents the value of the parameter, and each polyline represents one input configuration. The table below shows the configuration details.

the modellers' requests conveyed via email, we incorporated a line chart to depict the model outcomes. See Figure 6.7. The x-axis of the chart corresponds to the number of days since the first date in the Scottish data set, while the y-axis represents the population. Line chart and other classic visual designs are widely used by the modellers, they are familiar with these designs and can easily interpret the results. The line chart is coordinated with the parallel coordinates plot, enabling the modellers to select a subset of the input parameters and quickly identify the corresponding model outcomes. A focus+context technique is used to highlight the selected subset of the input parameters in the parallel coordinates plot.

#### Meeting #4, View #3 - Nov 2020

On 25 Nov 2020, the group convened for the fourth meeting, held just a day after the announcement of the gathering rules for Christmas in the UK. During the meeting, we received feedback from the RAMPVis team on the new view of the input parameters, a table with glyphs. See Figure 6.8. We incorporated this table view featuring

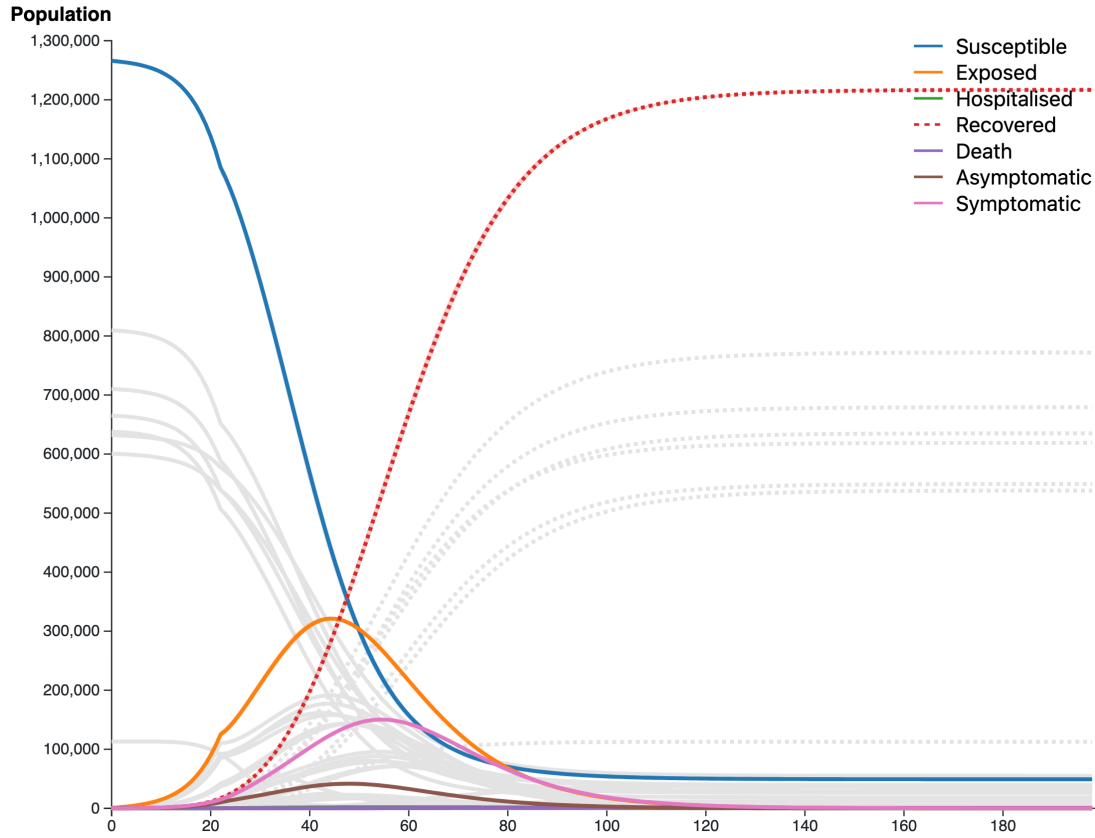


Figure 6.7: A line chart depicting the model outcomes. The x-axis of the chart corresponds to the number of days since the first date in the Scottish data set, while the y-axis represents the population. To differentiate between different population categories, a colormap was incorporated: `susceptible`, `exposed`, `hospitalised`, `recovered`, `death`, `asymptomatic`, and `symptomatic`. The focus+context technique is used here to highlight the outcome of the current configuration, while the grey lines represent other outcomes. On day 20, there is an unusual spike which was later identified as caused by an error in the model.

glyphs to depict all 160 input parameter configurations, following discussions with the modellers. Each row represents an input configuration, and each column represents an input parameter. The table view enables the modellers to sort configurations by individual input parameters. Each parameter value is symbolised by a bar glyph, the colour and length correspond to its deviation from the average value of 160 predictions.

The table view provides the functionality to sort the parameters according to their values and can be dynamically updated by brushing the parallel coordinates plot for the input parameters in Figure 6.6. The line chart in Figure 6.7 can be quickly updated to display the corresponding model outcomes by clicking on the configuration index in the table view.





Figure 6.8: The table view depicting all 160 input parameter configurations. The view enables the modellers to sort parameter values and identify interesting configurations. Each row represents an input configuration, and each column represents an input parameter. Upon clicking on a row, the line chart in [Figure 6.7](#) is updated to display the corresponding model outcomes. Clicking on the column header sorts the table by the parameter values.

### Meeting #5 - Dec 2020

On 9 Dec 2020, a week after the end of the second national lockdown in the UK, with England facing a stricter three-tier restriction policy, the group convened for the fifth meeting. At this point, we still had not met with the modellers, all communications and discussions took place via email. The RAMPVis team decided to organise a meeting with the modellers to present our prototype for feedback.

### Meeting #6, Views #4 & 5, Feedback #1 - Dec 2020

On 10 Dec 2020, we finally met with modellers from Durham University, the University of Edinburgh, the University of Exeter, the University of Glasgow, the London School of Hygiene & Tropical Medicine, for the first time. In contrast to sharing screenshots via email and deploying a website with a live view of our development (which they might not have been proficient in using), we delivered a live presentation, fielding numerous questions. The modellers were pleased with the dashboard, and a list of ad hoc requirements was provided. Furthermore, we collected insightful feedback that we elaborate on in detail in [Section 6.5](#).

1. The modellers found that the parallel coordinates plot is useful in identifying outliers, and requested the incorporation of another one for the model outcomes. Given that the outcome data mirrors the input in a multivariate format, employing a parallel coordinates plot could potentially be useful. We implemented this as shown in [Figure 6.9](#).
2. The modellers requested that all the simulation results be displayed in the line

chart, with the current one highlighted. This resembles their usual workflow for analysing multiple simulation outcomes. We implemented this as shown in [Figure 6.7](#).

3. The modellers requested the incorporation of a scatterplot to visualise the model outcomes, specifically a Principal Component Analysis (PCA) result obtained from another VIS volunteer team. The motivation behind this is to reduce the dimensionality and identify key parameters. We implemented this as shown in [Figure 6.10](#).
4. The modellers requested all views to be coordinated with each other, enabling observation of relationships between input parameters and model outcomes through interaction.
  - (a) Brushing on the input parallel coordinates plot ([Figure 6.6](#)) highlights the corresponding input configurations in both the table view ([Figure 6.8](#)) and scatterplot ([Figure 6.10](#)).
  - (b) Brushing on the scatterplot ([Figure 6.10](#)) for input configurations highlights the corresponding input configurations in both the table view ([Figure 6.8](#)) and input parallel coordinates plot ([Figure 6.6](#)).
  - (c) Clicking on a specific row in the table view ([Figure 6.8](#)) renders the corresponding output data in both the output parallel coordinates plot ([Figure 6.9](#)) and line chart ([Figure 6.7](#)).

Furthermore, we received the exciting news that initial funding had been successfully secured [291], which led to the transition of our volunteer work to a team of paid developers, who would continue with further implementation of the project.

### **Meeting #7, Feedback #2 - Mar 2021**

On 25 Mar 2021, the UK was in the process of cautiously lifting its third national lockdown, the ‘rule of two’ was still in place. The group convened for the seventh meeting, where we received further feedback from the modelling team on our implementation. We detail the feedback in [Section 6.5](#).

### **Last Commit - Apr 2021**

By 28 Apr 2021, more restrictive measures were abolished, although the prohibition

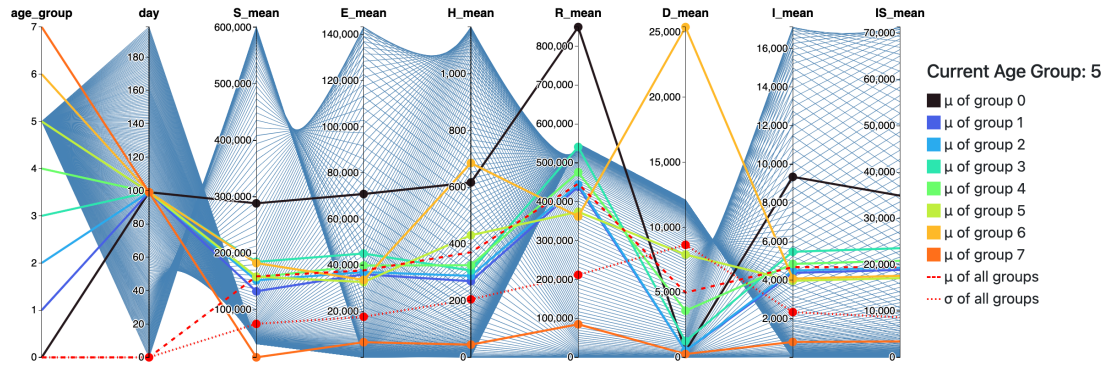


Figure 6.9: A parallel coordinates plot depicting the model outcomes by age group 5. As requested by the modellers, each blue line represents one simulation outcome, and each coloured line represents the age group’s mean. In addition, the dotted red line ... represents the group’s standard deviation, and the dashed red line --- represents the mean of all groups.

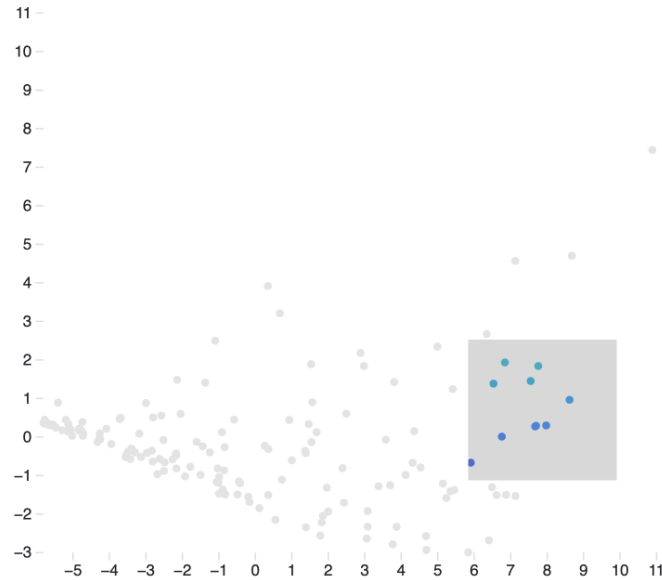


Figure 6.10: A scatterplot depicting the PCA outcome from another VIS volunteer group, was added upon request by the modellers. Upon brushing, the selected configurations are highlighted in the table view in [Figure 6.8](#).

on mixing between households was still in effect. On this day, we made our last commit to our GitHub repository. This act signified the completion of our volunteer work, as we had smoothly transitioned all tasks to a team of paid developers.

During the entire development process, our meetings were conducted exclusively online, and our communication relied heavily on email correspondence. Despite the lack of in person interactions, we successfully met the initial requirements of the modellers and delivered a VIS solution that received very positive feedback from the SCRC modelling team.

## **Meeting #8, Feedback #3 - May 2021**

On 19 May 2021, the UK was viewing the light at the end of the Covid tunnel, weddings and funerals were still restricted to 30 people, and indoor gatherings of more than two households were still banned. The group convened for the eighth and final volunteer meeting. During this final meeting, a modeller joined and gave us some in-depth feedback on the influence our work had on their modelling process, as well as suggesting potential improvements. We detail the feedback in [Section 6.5](#).

## **6.5 Domain Expert Feedback**

In this section, we share the invaluable feedback collected from the modellers. Meeting #6 and #7 were held prior to the conclusion of our development, serving as an iterative process of refinement aimed at validating and improving our visual designs while ensuring their relevance and utility to domain experts. Meeting #8 was held after the conclusion of our development, functioning as a means to gather feedback on our work and to identify potential future work. Three domain experts in statistics from Durham University, the University of Exeter, and the University of Glasgow, were invited to join these meetings.

### **6.5.1 Summary of Feedback**

In this section, we provide a summary of the feedback collected from the domain experts during our meetings.

#### **Appreciation for Interaction and Visualisation Design**

The experts commended the visual designs for effectively depicting the relative importance of input parameters on model predictions, highlighting the utility of interactive graphics in understanding the significance of different inputs. The ability to visually present the connection between input and output parameters was particularly appreciated, emphasising the value of visual techniques in elucidating the relationships between variables.

#### **Identification of Ineffective Parameter Combinations**

The linked visual designs were recognised for their potential to help identify ineffective parameter combinations, aiding in the optimisation and calibration process by revealing which combinations may not be useful. The ability to filter out redundant input parameter configurations was seen as beneficial for focusing on the most influential parameters, thereby reducing the dimensionality of the problem.

### **Potential for Identifying Model Discrepancy**

There was interest in the potential of visual designs to aid in identifying model discrepancies when observational data becomes available, highlighting the importance of visualising observational data alongside model predictions.

### **Overview of Input Parameters and Distributions**

The table view was praised for providing a clear overview of input parameters and their distributions, enabling quick identification of influential parameters and possible adjustments, as well as the elimination of unnecessary complexities.

## **6.5.2 Detailed Feedback**

In this section, we present some of the original quotes collected from domain experts during these meetings.

### **Domain Expert #1 - Professor in Statistics, Durham University**

On 25 March 2021, Meetings #6 and #7, we presented the dashboard through screen-sharing demonstrations, the domain expert appreciated the interactions provided by the visual designs in depicting the relative importance of different input parameters on the model's predictions. *“The visualisations are able to show how important a particular input is for a particular output.”*

In addition, the ability to visually present the link between the input and output parameters. *“The real interesting game here is the connection techniques to understand the relations between input and output.”*

The linked visual designs also potentially enable the domain expert to identify ineffective parameter combinations. *“The different configurations is the sort of history of calibration and by looking at those visualisations you can start saying certain combinations may not be useful.”*

The inclusion of a PCA plot was seen as a significant step towards dealing with feature selection. The expert suggested adding two further plots depicting, MDS and possibly ICA, to support model calibration using history matching. *“to perform history matching MDS is really what we use. The PCA plot is already very informative ... t-sne like methods are ill suited for the task.”*

Furthermore, the domain expert also expressed interest in the potential of our visual designs to aid in identifying model discrepancy, when the observational data becomes available. *“The visualisation would be helpful in identifying model discrepancies when we eventually plot the observational data.”*

### **Domain Expert #2 - Professor in Statistics, the University of Exeter**

On the same date, during Meeting #6 and #7, the domain expert was pleased with the ability of the visual designs to provide the potential to filter redundant input parameter configurations, enabling users to concentrate on the most influential configurations. *“For particular input configurations after filtering, the visualisation shows that some of the input parameters can be ignored, which reduces the dimensionality of the problem, and we can focus on the important parameters.”*

The domain expert also noted the usefulness of the PCA plot and suggested to replace the method with MPCA [245] to further support the process of detecting implausible input values *“One approach is to look for inputs configurations which would produce implausible outputs, we work with a sort of implausibility statistical measure”*.

### **Domain Expert #3 - Assistant Professor in Statistics, the University of Glasgow**

On 19 May 2021, Meeting #8, the domain expert praised the visual designs’ ability to provide a clear overview of the input parameters and their distributions. This enables them to quickly identify possible adjustments they can make to their input parameters, as well as to identify the most influential parameters. *“The table view is really useful in showing how close those input parameters are to the threshold, which is very useful to understand affordability.”*

The domain expert also noted that some overlapping distributions can be ruled out quickly via the interactivity provided by our visual designs, this enables them to

eliminate unnecessary complexities and increase the overall efficiency of their model. *“It’s fairly obvious that some of the parameters can be ruled out quite quickly, including some overlapping distributions.”*

An avenue for future work, as unanimously identified by all three domain experts, involves integrating new visual designs to render and compare observational data against model predictions effectively.

## 6.6 Limitations

Due to the impact of the pandemic, the project was conducted in a fully virtual manner, with all meetings and discussions taking place online, between a large group of researchers from different disciplines and different institutions. In total, 33 VIS researchers and 7 modellers were involved in this volunteer work. The development was ad hoc in some ways due to the unprecedented nature of the pandemic. This resulted in a number of limitations, which we will discuss in this section.

**Lack of Novel and Advanced Visual Designs:** Operating under a time constraint, the primary objective of our project centered on offering immediate visual analysis assistance to the modellers. Thus, we were unable to explore the inclusion of innovative and advanced visual design approaches. Instead, we integrated a series of classic views, such as line charts and scatterplots. These are visual designs commonly leveraged by modellers in their day-to-day research. Interestingly, the modellers welcomed the introduction of a less conventional (to them) visualisation technique: parallel coordinates. They had never previously employed this technique, and its introduction proved beneficial to their research. Consequently, they expressed a desire for the incorporation of an additional parallel coordinates to assist in the visualisation of model outcomes.

We believe that this is a testament to the effectiveness of advanced visual designs in enhancing the modellers’ understanding of their models, this signals the possibility for future inclusion of more sophisticated visual designs.

**Lack of Formal Requirements Gathering:** We were unable to meet with the modellers until a particularly late stage. Instead, we had to rely on email correspondence, which was arguably not as effective as face-to-face or even virtual meetings. In a tra-

ditional software engineering project, requirements are gathered through a series of meetings and discussions with end users. This did not occur in our case.

This resulted in a lack of proper requirement gathering, which in turn led to a number of challenges during the development process. For example, the modellers made ad hoc requests to incorporate different views at different stages of the project, resulting in unexpected changes on the development side. This could have been avoided if we had a better understanding of their requirements from the beginning.

**Dynamic Group Membership:** The group membership was dynamic, with researchers joining and leaving the group at different stages of the project. This introduced some lack of continuity, as newcomers had to spend time to familiarise themselves with the project. Furthermore, members came from different disciplines, with different levels of expertise in visualisation. This has resulted in a lack of consistency in the development process, as different members have different ideas on how to implement the views. The responsibility of each member, apart from the only developer in the group, was not clearly defined.

**Uncertain Project Direction:** The exact direction of the project was not clearly defined from the outset. Numerous details remained unknown to us during the development process, such as the exact purpose of the visualisation, the target audience, and the end product. Consequently, the final product suffered from suboptimal utilisation of screen space, as additional views were requested, the implementation of a multiview display design or collapsible views became time-constrained and unattainable.

**Other Technical Limitations:** Some additional technical limitations include:

- Real-time updating: Coupling the simulation with the visual rendering directly would have been very beneficial to the project, e.g., computational steering.
- Standardisation: Standardisation of the data format would be beneficial to all participants.
- Interpretability: A more formal evaluation of how interpretable our visual representations are would be beneficial, e.g., presenting complex epidemiological concepts in a clear and understandable manner to a wider audience.



## 6.7 Conclusions

In this chapter, we present the stories behind the development of EnsembleDashVis, an interactive dashboard designed to visualise the input parameters and outcomes of an ABC-SMC inference model used to analyse COVID-19 data collected during the first wave of the outbreak in Scotland.

Given the multitude of uncertainties and challenges during this exceptional period, a considerable amount of information was unavailable to us during the development process. It was only through the Scottish COVID-19 Response Consortium Stakeholder Report [283], published in late 2021, and various publications [301, 302, 305, 306] that unveiled the remarkable endeavours undertaken by other volunteer teams, gained additional insight and details.

While this chapter distincts from the primary focus on EHR Vis, its challenges and solutions align closely with the broader EHR Vis framework. The visualisation of pandemic simulations, like EHR data, requires handling large, multidimensional data sets, supporting interactive exploration, and ensuring that insights remain interpretable for diverse stakeholders. The iterative, expert-driven design process in this chapter mirrors the methodological approach of previous chapters, emphasising the importance of domain collaboration in visualisation development.



## Chapter 7

## Conclusion

*“Never believe that one number on its own can be meaningful.”*

– Hans G. Rosling, Physician and Statistician (1948 - 2017)

This chapter concludes the discussions presented in this thesis, synthesising key findings and outlining future research avenues in EHR Vis. The research conducted has significantly advanced the field by addressing fundamental challenges in textual, spatial, and temporal EHR data visualisation.

## 7.1 Key Contributions and Findings

This thesis contributes to EHR Vis by developing novel techniques that improve the interpretability, scalability, and usability of complex EHR data. The research follows an iterative, expert-driven design approach, ensuring that the proposed visualisations align with real-world healthcare needs.

Unlike previous works that focus on a single visualisation problem, this thesis takes a multidimensional approach. It integrates textual, spatial, and temporal data representations to support clinical decision-making, where insights often emerge from the interplay between different data modalities.

The findings extend beyond EHRs and can be applied to any domain dealing with complex and multimodal data. The emphasis on human-centred design and interdisciplinary collaboration provides a transferable model that can guide future research in other fields.

### 7.1.1 Summary of Chapters

This thesis begins with [Chapter 1](#), which introduces the research landscape of EHR Vis, discussing the challenges of visualising complex patient records, the motivation behind this research, and the methodology employed. It provides an overview of data visualisation techniques, their relevance to healthcare, and the specific gaps in current EHR visualisation approaches that this thesis seeks to address. The chapter also outlines the research contributions, highlighting the significance of interactive, scalable, and domain-expert-informed visualisation tools that serve as a critical reference for subsequent chapters.

[Chapter 2](#) presents a state-of-the-art review of interactive EHR Vis techniques, systematically classifying existing research and identifying gaps in the field. By analysing a broad spectrum of literature, this chapter establishes a taxonomy of EHR Vis methodologies, categorising techniques based on their visualisation approach, target data type, and intended user interaction. The findings reveal the need for novel solutions that address the challenges of heterogeneous EHR data, informing the subsequent chapters' research directions.

In [Chapter 3](#), the focus shifts to text data visualisation with the introduction of LetterVis, a novel tool for structuring and analysing unstructured clinical text. LetterVis enables clinicians to explore medical letters through an interactive, letter-space representation, allowing for rapid identification of key information such as medications, diagnoses, and treatment histories. The chapter discusses the iterative development process conducted with domain experts, ensuring that the final visualisation effectively supports clinical decision-making. Evaluations demonstrate its ability to enhance readability and facilitate comparative analysis of multiple documents.

[Chapter 4](#) introduces a spatial data visualisation technique through the development of a novel hybrid cartogram algorithm for Demers Cartograms. This approach incorporates river networks to enhance the spatial accuracy and legibility of geospatial health data. The proposed solution improves existing cartogram techniques by preserving geographic relationships while ensuring that distortions introduced in the visualisation process do not hinder interpretability. A user study validates the effectiveness of this method, showing its potential for applications in epidemiological mapping and healthcare resource planning.

Building on the need for effective time series visualisation, [Chapter 5](#) presents Time Series Map, a scalable approach for organising and visualising temporal patterns in long-term health data. The chapter outlines the design of a hierarchical visualisation system that allows clinicians to extract meaningful insights from extensive patient histories. By structuring time series events into visual hierarchies, this method supports efficient exploration and pattern recognition, particularly for chronic disease monitoring. User evaluations highlight its usefulness in summarising long-term trends and identifying irregularities in patient data.

[Chapter 6](#) offers a retrospective analysis of EnsembleDashVis, a visualisation dashboard developed for COVID-19 forecast simulations. This project was a collaborative effort that involved more than 40 experts in various fields, including epidemiology, mathematics, and data science. The chapter details the design and deployment of the dashboard, which was used to analyse epidemiological models during the pandemic. It also reflects on the challenges of remote interdisciplinary collaboration, emphasising the importance of agile development and expert-driven design in visualisation research.

## 7.2 Future Work and Challenges

While this research has contributed significantly to the advancement of EHR Vis, several challenges remain for future exploration. One key area is the integration of visualisation modalities, where a unified system could allow clinicians to seamlessly transition between textual, spatial, and temporal views. Such an approach would enable a more comprehensive analysis of patient records, reducing cognitive load and improving decision-making efficiency.

Since the publication of EHR STAR [311], the field of EHR Vis has gained increasing attention from researchers across various disciplines. Rapid advancements in interactive EHR Vis systems and techniques since then suggest that a new and updated survey may be necessary to capture the latest developments and trends.

Another challenge lies in EHR data set accessibility and standardisation. The lack of high-quality open-access EHR data sets has historically impeded research progress. The mini-survey on open-access EHR data sets in Chapter 2 offers a starting point, but there is a clear need for a more comprehensive review of available data sets. Expanding this section into a full survey would significantly benefit the research community by saving researchers valuable time and streamlining their work.

The role of AI-assisted and automated visualisation also presents an exciting avenue for further research. Machine learning techniques could be integrated into visualisation systems to identify patterns, highlight anomalies, and provide predictive insights. Exploring how AIs can be effectively combined with human-centric visualisation techniques remains an open challenge.

Enhancing interactive and multimodal exploration is another potential direction. While this thesis introduces novel visualisation tools, future work should focus on improving interactivity, allowing users to query, manipulate, and cross-analyse data across different visual representations. This could lead to more intuitive and powerful decision support systems for clinicians.

Finally, user-centred evaluation and deployment in real-world clinical environments is essential. While this thesis incorporates domain expert feedback, broader clinical validation is required to assess the practical impact of these visualisation tools. The deployment of them in healthcare settings and the study of their influence on clinical

workflows will be crucial in refining and adapting them to widespread adoption.

In conclusion, this thesis advances the field of EHR Vis by addressing key challenges in text, geospatial, and temporal data visualisation. The proposed techniques provide novel and innovative solutions that improve the usability and interpretability of complex medical records. By emphasising human-centred design, interdisciplinary collaboration, and scalability, this thesis establishes a foundation for future advancements in interactive EHR Vis. As the field evolves, continued integration of AI, multimodal visualisation, and real-world clinical validation will be critical in shaping the next generation of EHR Vis systems.





# Bibliography

- [1] William Playfair. *The Commercial and Political Atlas: Representing, by Means of Stained Copper-plate Charts, the Progress of the Commerce, Revenues, Expenditure and Debts of England During the Whole of the Eighteenth Century*. T. Burton, 1801.
- [2] Edward Heawood. “John Adams and His Map of England”. In: *The Geographical Journal* 79.1 (1932), pp. 37–44. ISSN: 0016-7398. DOI: [10.2307/1784518](https://doi.org/10.2307/1784518). JSTOR: [1784518](https://www.jstor.org/stable/1784518). (Visited on 07/03/2022).
- [3] Erwin Raisz. “The Rectangular Statistical Cartogram”. In: *Geographical Review* 24.2 (Apr. 1934), p. 292. ISSN: 00167428. DOI: [10.2307/208794](https://doi.org/10.2307/208794). JSTOR: [208794](https://www.jstor.org/stable/208794). (Visited on 01/15/2022).
- [4] S. E. Robertson and K. Sparck Jones. “Relevance Weighting of Search Terms”. In: *Journal of the American Society for Information Science* 27.3 (May 1976), pp. 129–146. ISSN: 00028231. DOI: [10.1002/asi.4630270302](https://doi.org/10.1002/asi.4630270302).
- [5] Rebecca Wirfs-Brock, Brian Wilkerson, and Lauren Wiener. *Designing Object-Oriented Software*. Englewood Cliffs, N.J: Prentice Hall, 1990. ISBN: 978-0-13-629825-0.
- [6] B. Johnson and B. Shneiderman. “Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures”. In: *Proceeding Visualization '91*. San Diego, CA, USA: IEEE Comput. Soc. Press, 1991, pp. 284–291. ISBN: 978-0-8186-2245-8. DOI: [10.1109/VISUAL.1991.175815](https://doi.org/10.1109/VISUAL.1991.175815). (Visited on 10/07/2023).
- [7] Ben Shneiderman. “Tree Visualization with Tree-Maps: 2-d Space-Filling Approach”. In: *ACM Transactions on Graphics* 11.1 (Jan. 1992), pp. 92–99. ISSN: 0730-0301, 1557-7368. DOI: [10.1145/102377.115768](https://doi.org/10.1145/102377.115768). (Visited on 07/10/2022).
- [8] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. “LifeLines: Visualizing Personal Histories”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Common Ground*. New York, New York, USA: ACM Press, 1996, 221–ff. ISBN: 0-89791-777-4. DOI: [10.1145/238386.238493](https://doi.org/10.1145/238386.238493).
- [9] B. Shneiderman. “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”. In: *Proceedings 1996 IEEE Symposium on Visual Languages*. Boulder, CO, USA: IEEE Comput. Soc. Press, 1996, pp. 336–343. ISBN: 978-0-8186-7508-9. DOI: [10.1109/VL.1996.545307](https://doi.org/10.1109/VL.1996.545307). (Visited on 02/24/2024).
- [10] Ilias Iakovidis. “Towards Personal Health Record: Current Situation, Obstacles and Trends in Implementation of Electronic Healthcare Record in Europe”. In: *International Journal of Medical Informatics* 52.1-3 (1998), pp. 105–115. ISSN: 13865056. DOI: [10.1016/s1386-5056\(98\)00129-4](https://doi.org/10.1016/s1386-5056(98)00129-4).

- [11] C Plaisant, R Mushlin, A Snyder, J Li, D Heller, and B Shneiderman. “LifeLines: Using Visualization to Enhance Navigation and Analysis of Patient Records.” In: *Proceedings. AMIA Symposium* 48.9 (Sept. 1998), pp. 76–80. ISSN: 1531-605X. DOI: [10.1016/B978-155860915-0/50038-X](https://doi.org/10.1016/B978-155860915-0/50038-X).
- [12] J. Bertin. “Graphics and Graphic Information Processing”. In: *Readings in Information Visualization*. Ed. by Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. Morgan Kaufmann, 1999, pp. 62–65. ISBN: 1-55860-533-9.
- [13] Stuart K. Card, Jock D. Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision to Think*. The Morgan Kaufmann Series in Interactive Technologies. San Francisco, Calif: Morgan Kaufmann Publishers, 1999. ISBN: 978-1-55860-533-6.
- [14] C Friedman and G Hripcsak. “Natural Language Processing and Its Future in Medicine”. In: *Academic Medicine* 74.8 (Aug. 1999), pp. 890–5. ISSN: 1040-2446. DOI: [10.1097/00001888-199908000-00012](https://doi.org/10.1097/00001888-199908000-00012).
- [15] Ying-Huey Fua, M.O. Ward, and E.A. Rundensteiner. “Hierarchical Parallel Coordinates for Exploration of Large Datasets”. In: *Proceedings Visualization '99 (Cat. No.99CB37067)*. San Francisco, CA, USA: IEEE, 1999, pp. 43–508. ISBN: 978-0-7803-5897-3. DOI: [10.1109 / VISUAL.1999.809866](https://doi.org/10.1109/VISUAL.1999.809866). (Visited on 08/12/2021).
- [16] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. “PhysioBank, PhysioToolkit, and PhysioNet”. In: *Circulation* 101.23 (June 2000). ISSN: 0009-7322. DOI: [10.1161/01.CIR.101.23.e215](https://doi.org/10.1161/01.CIR.101.23.e215).
- [17] S. Havre, B. Hetzler, and L. Nowell. “ThemeRiver: Visualizing Theme Changes over Time”. In: *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings. IEEE Comput. Soc, 2000*, pp. 115–123. ISBN: 0-7695-0804-9. DOI: [10.1109/INFVIS.2000.885098](https://doi.org/10.1109/INFVIS.2000.885098).
- [18] John Stasko and Eugene Zhang. “Focus+context Display and Navigation Techniques for Enhancing Radial, Space-Filling Hierarchy Visualizations”. In: *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings. IEEE Comput. Soc, 2000*, pp. 57–65. ISBN: 0-7695-0804-9. DOI: [10.1109/INFVIS.2000.885091](https://doi.org/10.1109/INFVIS.2000.885091).
- [19] W. Horn, C. Popow, and L. Unterasinger. “Support for Fast Comprehension of ICU Data: Visualization Using Metaphor Graphics”. In: *Methods of Information in Medicine* (2001). ISSN: 00261270. DOI: [10.1055/s-0038-1634202](https://doi.org/10.1055/s-0038-1634202).
- [20] Ulrich John, Elke Hensel, Lü, Jan Demann, Marion Piek, Sybille Sauer, Christiane Adam, Gabriele Born, Dietrich Alte, Eberhard Greiser, Ursula Haertel, Hans-Werner Hense, Johannes Haerting, Stefan Willich, and Christof Kessler. “Study of Health in Pomerania (SHIP): A Health Examination Survey in an East German Region: Objectives and Design”. In: *Sozial- und Präventivmedizin SPM* 46.3 (May 2001), pp. 186–194. ISSN: 0303-8408. DOI: [10.1007/BF01324255](https://doi.org/10.1007/BF01324255).
- [21] Robert Kosara and Silvia Miksch. “Metaphors of Movement: A Visualization and User Interface for Time-Oriented, Skeletal Plans”. In: *Artificial Intelligence in Medicine* 22.2 (May 2001), pp. 111–131. ISSN: 09333657. DOI: [10.1016/S0933-3657\(00\)00103-2](https://doi.org/10.1016/S0933-3657(00)00103-2).
- [22] Edward R. Tufte. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, Conn: Graphics Press, 2001. ISBN: 978-0-9613921-4-7.

- [23] M. Weber, M. Alexa, and W. Muller. “Visualizing Time-Series on Spirals”. In: IEEE Symposium on Information Visualization, 2001. INFOVIS 2001. IEEE, 2001, pp. 7–13. ISBN: 0-7695-7342-5. DOI: [10.1109/INFVIS.2001.963273](https://doi.org/10.1109/INFVIS.2001.963273).
- [24] P.J Atkinson and D.J Unwin. “Density and Local Attribute Estimation of an Infectious Disease Using MapInfo”. In: *Computers & Geosciences* 28.9 (Nov. 2002), pp. 1095–1105. ISSN: 00983004. DOI: [10.1016/S0098-3004\(02\)00026-2](https://doi.org/10.1016/S0098-3004(02)00026-2).
- [25] Bortins Ian, Demers Steve, and Clarke Keith. *Cartogram Home*. <http://www.ncgia.ucsb.edu/projects/Cartogram.Central/index.html>. 2002. (Visited on 03/03/2022).
- [26] D.A. Keim. “Information Visualization and Visual Data Mining”. In: *IEEE Transactions on Visualization and Computer Graphics* 8.1 (2002), pp. 1–8. DOI: [10.1109/2945.981847](https://doi.org/10.1109/2945.981847).
- [27] Luca Chittaro, Carlo Combi, and Giampaolo Trapasso. “Data Mining on Temporal Data: A Visual Approach and Its Clinical Application to Hemodialysis”. In: *Journal of Visual Languages & Computing* 14.6 (Dec. 2003), pp. 591–620. ISSN: 1045926X. DOI: [10.1016/j.jvlc.2003.06.003](https://doi.org/10.1016/j.jvlc.2003.06.003).
- [28] Mark Harrower and Cynthia A. Brewer. “ColorBrewer.Org: An Online Tool for Selecting Colour Schemes for Maps”. In: *The Cartographic Journal* 40.1 (June 2003), pp. 27–37. ISSN: 0008-7041, 1743-2774. DOI: [10.1179/000870403235002042](https://doi.org/10.1179/000870403235002042). (Visited on 05/25/2024).
- [29] E.D Liddy. “Natural Language Processing”. In: *Encyclopedia of Library and Information Science*. Ed. by Miriam A. Drake. 2nd ed. New York: Marcel Dekker, 2003. ISBN: 978-0-8247-2078-0.
- [30] Ragnar Bade, Stefan Schlechtweg, and Silvia Miksch. “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*. May 2004. New York, New York, USA: ACM Press, 2004, pp. 105–112. ISBN: 1-58113-702-8. DOI: [10.1145/985692.985706](https://doi.org/10.1145/985692.985706).
- [31] Michael Balzer, Andreas Noack, Oliver Deussen, and Claus Lewerentz. “Software Landscapes: Visualizing the Structure of Large Software Systems”. In: *Eurographics / IEEE VGTC Symposium on Visualization* (2004), 6 pages. ISSN: 1727-5296. DOI: [10.2312/VISSYM/VISSYM04/261-266](https://doi.org/10.2312/VISSYM/VISSYM04/261-266). (Visited on 07/10/2022).
- [32] Olivier Bodenreider. “The Unified Medical Language System (UMLS): Integrating Biomedical Terminology”. In: *Nucleic Acids Research* 32.DATABASE ISS. (Jan. 2004), pp. 267D–270. ISSN: 03051048. DOI: [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061).
- [33] Dina Goren-Bar, Yuval Shahr, Maya Galperin-Aizenberg, David Boaz, and Gil Tahan. “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*. ACM Press, 2004, p. 171. ISBN: 1-58113-867-9. DOI: [10.1145/989863.989889](https://doi.org/10.1145/989863.989889).
- [34] *IEEE Visualization 2004 Contest*. <http://vis.computer.org/vis2004contest/>. 2004.
- [35] Pak Chung Wong and J. Thomas. “Visual Analytics”. In: *IEEE Computer Graphics and Applications* 24.5 (Sept. 2004), pp. 20–21. ISSN: 0272-1716, 1558-1756. DOI: [10.1109/MCG.2004.39](https://doi.org/10.1109/MCG.2004.39). (Visited on 07/17/2022).

- [36] Marc van Kreveld and Bettina Speckmann. “On Rectangular Cartograms”. In: *Algorithms – ESA 2004*. Ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Susanne Albers, and Tomasz Radzik. Vol. 3221. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 724–735. ISBN: 978-3-540-23025-0 978-3-540-30140-0. DOI: [10.1007/978-3-540-30140-0.64](https://doi.org/10.1007/978-3-540-30140-0.64). (Visited on 01/02/2022).
- [37] Dominique Brodbeck, Roland Gasser, and Markus Degen. “Enabling Large-Scale Telemedical Disease Management through Interactive Visualization”. In: *Proceedings of MIE 2005*. Vol. 1. Geneva: European Notes in Medical Informatics, 2005, pp. 1172–1177.
- [38] Tracy D Gunter and Nicolas P Terry. “The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions”. In: *Journal of Medical Internet Research* 7.1 (Mar. 2005), e3. ISSN: 1438-8871. DOI: [10.2196/jmir.7.1.e3](https://doi.org/10.2196/jmir.7.1.e3). (Visited on 11/24/2020).
- [39] Klaus Hinum, Silvia Miksch, Wolfgang Aigner, Susanne Ohmann, Christian Popow, Margit Pohl, and Markus Rester. “Gravi++: Interactive Information Visualization to Explore Highly Structured Temporal Data”. In: *Journal of Universal Computer Science* 11.11 (2005), pp. 1792–1805. ISSN: 0958695X. DOI: [10.3217/jucs-011-11-1792](https://doi.org/10.3217/jucs-011-11-1792).
- [40] James J. Thomas and Kristin A. Cook, eds. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005. ISBN: 0-7695-2323-4.
- [41] R. Xu and D. WunschII. “Survey of Clustering Algorithms”. In: *IEEE Transactions on Neural Networks* 16.3 (May 2005), pp. 645–678. DOI: [10.1109/TNN.2005.845141](https://doi.org/10.1109/TNN.2005.845141).
- [42] Wolfgang Aigner and Silvia Miksch. “CareVis: Integrated Visualization of Computerized Protocols and Temporal Patient Data”. In: *Artificial Intelligence in Medicine* 37.3 (2006), pp. 203–218. ISSN: 09333657. DOI: [10.1016/j.artmed.2006.04.002](https://doi.org/10.1016/j.artmed.2006.04.002).
- [43] Jesse D. Blanton, Arie Manangan, Jamie Manangan, Cathleen A. Hanlon, Dennis Slate, and Charles E. Rupprecht. “Development of a GIS-based, Real-Time Internet Mapping Tool for Rabies Surveillance”. In: *International Journal of Health Geographics* 5 (2006), pp. 1–8. ISSN: 1476072X. DOI: [10.1186/1476-072X-5-47](https://doi.org/10.1186/1476-072X-5-47).
- [44] Tim Dwyer, Kim Marriott, and Peter J. Stuckey. “Fast Node Overlap Removal”. In: *Graph Drawing*. Ed. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Patrick Healy, and Nikola S. Nikolov. Vol. 3843. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 153–164. ISBN: 978-3-540-31425-7 978-3-540-31667-1. DOI: [10.1007/11618058.15](https://doi.org/10.1007/11618058.15). (Visited on 06/17/2021).
- [45] Jerry Fails, Amy Karlson, Layla Shahamat, and Ben Shneiderman. “A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across Multiple Histories”. In: *2006 IEEE Symposium On Visual Analytics And Technology*. IEEE, Oct. 2006, pp. 167–174. ISBN: 1-4244-0591-2. DOI: [10.1109/VAS.T.2006.261421](https://doi.org/10.1109/VAS.T.2006.261421).

- [46] Robert S Laramee and Robert Kosara. “Challenges and Unsolved Problems”. In: *Human-Centered Visualization Environments*. Vol. 4417. Springer-Verlag, 2006, pp. 231–254.
- [47] P. C. Tang, J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands. “Personal Health Records: Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption”. In: *Journal of the American Medical Informatics Association* 13.2 (Mar. 2006), pp. 121–126. ISSN: 1067-5027. DOI: [10.1197/jamia.M2025](https://doi.org/10.1197/jamia.M2025).
- [48] Alex A. T. Bui, Denise R. Aberle, and Hooshang Kangarloo. “TimeLine: Visualizing Integrated Patient Records”. In: *IEEE Transactions on Information Technology in Biomedicine* 11.4 (July 2007), pp. 462–473. ISSN: 1089-7771. DOI: [10.1109/TITB.2006.884365](https://doi.org/10.1109/TITB.2006.884365).
- [49] Fabricio A.B. Da Silva, Henrique F. Gagliardi, Eduardo Gallo, Maria A. Madope, Virgilio C. Neto, Ivan T. Pisa, and Domingos Alves. “IntegraEPI: A Grid-Based Epidemic Surveillance System”. In: *Studies in Health Technology and Informatics* 126. February (2007), pp. 197–206. ISSN: 18798365.
- [50] Bill Elliott. “Anything Is Possible: Managing Feature Creep in an Innovation Rich Environment”. In: *2007 IEEE International Engineering Management Conference*. Lost Pines, TX, USA: IEEE, July 2007, pp. 304–307. ISBN: 978-1-4244-2145-9. DOI: [10.1109/IEMC.2007.5235049](https://doi.org/10.1109/IEMC.2007.5235049). (Visited on 03/15/2024).
- [51] D. Guo. “Visual Analytics of Spatial Interaction Patterns for Pandemic Decision Support”. In: *International Journal of Geographical Information Science* 21.8 (Sept. 2007), pp. 859–877. ISSN: 1365-8816. DOI: [10.1080/13658810701349037](https://doi.org/10.1080/13658810701349037).
- [52] Paul Jen-Hwa Hu, Daniel Zeng, Hsinchun Chen, Catherine Larson, Wei Chang, Chunju Tseng, and James Ma. “System for Infectious Disease Information Sharing and Analysis: Design and Evaluation”. In: *IEEE Transactions on Information Technology in Biomedicine* 11.4 (July 2007), pp. 483–492. DOI: [10.1109/TITB.2007.893286](https://doi.org/10.1109/TITB.2007.893286).
- [53] David S. Pieczkiewicz, Stanley M. Finkelstein, and Marshall I. Hertz. “Design and Evaluation of a Web-Based Interactive Visualization System for Lung Transplant Home Monitoring Data.” In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* (Oct. 2007), pp. 598–602. ISSN: 1942-597X.
- [54] Sheng Gao, Darka Mioc, Francois Anton, Xiaolun Yi, and David J. Coleman. “Online GIS Services for Mapping and Sharing Disease Information”. In: *International Journal of Health Geographics* 7.1 (2008), p. 8. ISSN: 1476-072X. DOI: [10.1186/1476-072X-7-8](https://doi.org/10.1186/1476-072X-7-8).
- [55] Catalina Hallett. “Multi-Modal Presentation of Medical Histories”. In: *Proceedings of the 13th International Conference on Intelligent User Interfaces - IUI '08*. New York, New York, USA: ACM Press, 2008, p. 80. ISBN: 978-1-59593-987-6. DOI: [10.1145/1378773.1378785](https://doi.org/10.1145/1378773.1378785).
- [56] Janet L. Heitgerd, Andrew L. Dent, Kimberlee A. Elmore, Brian Kaplan, James B. Holt, Marilyn M. Metzler, Koren Melfi, Jennifer M. Stanley, Keisher Highsmith, Norma Kanarek, and Karen Frederickson Comer. “Community Health Status Indicators: Adding a Geospatial Component.” In: *Preventing chronic disease* 5.3 (July 2008), A96. ISSN: 1545-1151.



- [57] Petra Isenberg, Torre Zuk, Christopher Collins, and Sheelagh Carpendale. “Grounded Evaluation of Information Visualizations”. In: *Proceedings of the 2008 Conference on BEyond Time and Errors Novel evaluation Methods for Information Visualization - BELIV '08*. ACM Press, 2008, p. 1. DOI: [10.1145/1377966.1377974](https://doi.org/10.1145/1377966.1377974).
- [58] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. “Visual Analytics: Definition, Process, and Challenges”. In: *Information Visualization*. Ed. by Andreas Kerren, John T. Stasko, Jean-Daniel Fekete, and Chris North. Vol. 4950. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 154–175. ISBN: 978-3-540-70955-8 978-3-540-70956-5. DOI: [10.1007/978-3-540-70956-5\\_7](https://doi.org/10.1007/978-3-540-70956-5_7). (Visited on 07/17/2022).
- [59] Markus Reinhardt, Johannes Elias, Jürgen Albert, Matthias Frosch, Dag Harmesen, and Ulrich Vogel. “EpiScanGIS: An Online Geographic Surveillance System for Meningococcal Disease”. In: *International Journal of Health Geographics* 7.1 (2008), p. 33. ISSN: 1476-072X. DOI: [10.1186/1476-072X-7-33](https://doi.org/10.1186/1476-072X-7-33).
- [60] Ronald P. Stolk, Judith G. M. Rosmalen, Dirkje S. Postma, Rudolf A. de Boer, Gerjan Navis, Joris P. J. Slaets, Johan Ormel, and Bruce H. R. Wolffenbuttel. “Universal Risk Factors for Multifactorial Diseases”. In: *European Journal of Epidemiology* 23.1 (Jan. 2008), pp. 67–74. ISSN: 0393-2990. DOI: [10.1007/s10654-007-9204-4](https://doi.org/10.1007/s10654-007-9204-4).
- [61] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H. Stumpf. “Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems”. In: *Journal of The Royal Society Interface* 6.31 (July 2008), pp. 187–202. DOI: [10.1098/rsif.2008.0172](https://doi.org/10.1098/rsif.2008.0172). (Visited on 05/25/2023).
- [62] Barney Warf and Mort Winsberg. “The Geography of Religious Diversity in the United States”. In: *The Professional Geographer* 60.3 (Aug. 2008), pp. 413–424. ISSN: 0033-0124. DOI: [10.1080/00330120802046786](https://doi.org/10.1080/00330120802046786). (Visited on 12/08/2021).
- [63] Qian Yi, Richard E. Hoskins, Elizabeth A. Hillringhouse, Svend S. Sorensen, Mark W. Oberle, Sherrilynne S. Fuller, and James C. Wallace. “Integrating Open-Source Technologies to Build Low-Cost Information Systems for Improved Access to Public Health Data”. In: *International Journal of Health Geographics* 7.1 (2008), p. 29. ISSN: 1476-072X. DOI: [10.1186/1476-072X-7-29](https://doi.org/10.1186/1476-072X-7-29).
- [64] Vijayaraghavan Bashyam, William Hsu, Emily Watt, Alex A. T. Bui, Hooshang Kangarloo, and Ricky K. Taira. “Problem-Centric Organization and Visualization of Patient Imaging and Clinical Data”. In: *RadioGraphics* 29.2 (Mar. 2009), pp. 331–343. ISSN: 0271-5333. DOI: [10.1148/rg.292085098](https://doi.org/10.1148/rg.292085098).
- [65] Joseph Connors, Martin Krzywinski, Jacqueline Schein, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. “Circos : An Information Aesthetic for Comparative Genomics”. In: *Genome Research* 19.604 (2009), pp. 1639–1645. DOI: [10.1101/gr.092759.109.19](https://doi.org/10.1101/gr.092759.109.19).
- [66] Borden D. Dent, Jeffrey Torguson, and T. W. Hodler. *Cartography: Thematic Map Design*. 6th ed. New York: McGraw-Hill Higher Education, 2009. ISBN: 978-0-07-294382-5.

- [67] David V Ford, Kerina H Jones, Jean-Philippe Verplancke, Ronan A Lyons, Gareth John, Ginevra Brown, Caroline J Brooks, Simon Thompson, Owen Bodger, Tony Couch, and Ken Leake. “The SAIL Databank: Building a National Architecture for e-Health Research and Evaluation”. In: *BMC Health Services Research* 9.1 (Dec. 2009), p. 157. ISSN: 1472-6963. DOI: [10.1186/1472-6963-9-157](https://doi.org/10.1186/1472-6963-9-157).
- [68] Tamara Munzner. “A Nested Model for Visualization Design and Validation”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 921–928. ISSN: 1077-2626. DOI: [10.1109/TVCG.2009.111](https://doi.org/10.1109/TVCG.2009.111).
- [69] A. Slingsby, J. Dykes, and J. Wood. “Configuring Hierarchical Layouts to Address Research Questions”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 977–984. ISSN: 1077-2626. DOI: [10.1109/TVCG.2009.128](https://doi.org/10.1109/TVCG.2009.128). (Visited on 07/10/2022).
- [70] *The American Recovery and Reinvestment Act of 2009*. <https://www.congress.gov/bill/111th-congress/house-bill/1/text>. Jan. 2009.
- [71] T.D. Wang, Catherine Plaisant, Ben Shneiderman, Neil Spring, David Roseman, Greg Marchand, Vikramjit Mukherjee, and Mark Smith. “Temporal Summaries: Supporting Temporal Categorical Searching, Aggregation and Comparison”. In: *IEEE Transactions on Visualization and Computer Graphics* 15.6 (Nov. 2009), pp. 1049–1056. ISSN: 1077-2626. DOI: [10.1109/TVCG.2009.187](https://doi.org/10.1109/TVCG.2009.187).
- [72] Krist Wongsuphasawat and Ben Shneiderman. “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. ISBN: 978-1-4244-5283-5. DOI: [10.1109/VAST.2009.5332595](https://doi.org/10.1109/VAST.2009.5332595).
- [73] Ethem Alpaydin. *Introduction to Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press, 2010. ISBN: 978-0-262-01243-0.
- [74] Daniel J Friedman and R Gibson Parrish. “The Population Health Record: Concepts, Definition, Design, and Implementation”. In: *Journal of the American Medical Informatics Association* 17.4 (July 2010), pp. 359–366. ISSN: 1067-5027, 1527-974X. DOI: [10.1136/jamia.2009.001578](https://doi.org/10.1136/jamia.2009.001578). (Visited on 11/24/2020).
- [75] Michael-Rock Goldsmith, Thomas R. Transue, Daniel T. Chang, Rogelio Tornero-Velez, Michael S. Breen, and Curtis C. Dary. “PAVA: Physiological and Anatomical Visual Analytics for Mapping of Tissue-Specific Concentration and Time-Course Data”. In: *Journal of Pharmacokinetics and Pharmacodynamics* 37.3 (June 2010), pp. 277–287. ISSN: 1567-567X. DOI: [10.1007/s10928-010-9160-6](https://doi.org/10.1007/s10928-010-9160-6).
- [76] Denis Klimov, Yuval Shahar, and Meirav Taieb-Maimon. “Intelligent Selection and Retrieval of Multiple Time-Oriented Records”. In: *Journal of Intelligent Information Systems* 35.2 (Oct. 2010), pp. 261–300. ISSN: 0925-9902. DOI: [10.1007/s10844-009-0100-0](https://doi.org/10.1007/s10844-009-0100-0).
- [77] Natsuhiko Kumasaka, Yusuke Nakamura, and Naoyuki Kamatani. “The Textile Plot: A New Linkage Disequilibrium Display of Multiple-Single Nucleotide Polymorphism Genotype Data”. In: *PLoS ONE* 5.4 (2010), pp. 1–12. ISSN: 19326203. DOI: [10.1371/journal.pone.0010207](https://doi.org/10.1371/journal.pone.0010207).
- [78] Elena N. Naumova. “Visual Analytics for Immunologists: Data Compression and Fractal Distributions”. In: *Self/Nonself* 1.3 (July 2010), pp. 241–249. ISSN: 1938-2030. DOI: [10.4161/self.1.3.12876](https://doi.org/10.4161/self.1.3.12876).

- [79] Alexander Rind, Silvia Miksch, Wolfgang Aigner, Thomas Turic, and Margit Pohl. “VisuExplore: Gaining New Medical Insights from Visual Exploration”. In: *Proceedings of the 1st International Workshop on Interactive Systems in Healthcare (WISH@CHI2010)* (2010), pp. 149–152.
- [80] Francisco S. Roque, Laura Slaughter, and A Tkatšenko. “A Comparison of Several Key Information Visualization Systems for Secondary Use of Electronic Health Record Content”. In: *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents* June (2010), pp. 76–83.
- [81] Martijn D. Steenwijk, Julien Milles, Mark A. Buchem, Johan H.C. Reiber, and Charl P. Botha. “Integrated Visual Analysis for Heterogeneous Datasets in Cohort Studies”. In: *IEEE VisWeek Workshop on Visual Analytics in Health Care* (2010).
- [82] Hui Sun and Zhilin Li. “Effectiveness of Cartogram for the Representation of Spatial Data”. In: *The Cartographic Journal* 47.1 (Jan. 2010), pp. 12–21. ISSN: 0008-7041, 1743-2774. DOI: [10.1179/000870409X12525737905169](https://doi.org/10.1179/000870409X12525737905169). (Visited on 12/08/2021).
- [83] *VAST Challenge 2010 MC2 - Characterization of Pandemic Spread*. <https://old.datahub.io/dataset/vast-challenge-2010-mc2-characterization-of-pandemic-spread>. 2010.
- [84] *VAST Challenge 2010 MC3 - Tracing the Mutations of a Disease*. <https://old.datahub.io/dataset/vast-challenge-2010-mc3-tracing-the-mutations-of-a-disease>. 2010.
- [85] Brian Willison. “Advancing Meaningful Use: Simplifying Complex Clinical Metrics Through Visual Representation”. In: *Parsons Institute for Information Mapping (PIIM) Research* (2010).
- [86] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. “Survey of Visualization Techniques”. In: *Visualization of Time-Oriented Data*. London: Springer London, 2011, pp. 147–254. ISBN: 978-0-85729-078-6 978-0-85729-079-3. DOI: [10.1007/978-0-85729-079-3\\_7](https://doi.org/10.1007/978-0-85729-079-3_7). (Visited on 10/06/2023).
- [87] Daniel Dorling. “Area Cartograms: Their Use and Creation”. In: *The Map Reader*. Ed. by Martin Dodge, Rob Kitchin, and Chris Perkins. Chichester, UK: John Wiley & Sons, Ltd, Apr. 2011, pp. 252–260. ISBN: 978-0-470-97958-7 978-0-470-74283-9. DOI: [10.1002/9780470979587.ch33](https://doi.org/10.1002/9780470979587.ch33). (Visited on 12/05/2021).
- [88] Timothy Driscoll, Joseph L. Gabbard, Chunhong Mao, Oral Dalay, Maulik Shukla, Clark C. Freifeld, Anne Gatewood Hoen, John S. Brownstein, and Bruno W. Sobral. “Integration and Visualization of Host–Pathogen Data Related to Infectious Diseases”. In: *Bioinformatics (Oxford, England)* 27.16 (Aug. 2011), pp. 2279–2287. ISSN: 1460-2059. DOI: [10.1093/bioinformatics/btr391](https://doi.org/10.1093/bioinformatics/btr391).
- [89] Anthony Faiola and Chris Newlon. “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*. Vol. 6779 LNCS. 2011, pp. 119–128. ISBN: 978-3-642-21715-9. DOI: [10.1007/978-3-642-21716-6\\_13](https://doi.org/10.1007/978-3-642-21716-6_13).
- [90] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. “Visual Comparison for Information Visualization”. In: *Information Visualization* 10.4 (Oct. 2011), pp. 289–309. ISSN: 1473-8716. DOI: [10.1177/1473871611416549](https://doi.org/10.1177/1473871611416549).



- [91] David Gotz, Jimeng Sun, Nan Cao, and Shahram Ebadollahi. “Visual Cluster Analysis in Support of Clinical Decision Intelligence.” In: *AMIA ... Annual Symposium proceedings / AMIA Symposium*. *AMIA Symposium* (2011), pp. 481–490. ISSN: 1942597X.
- [92] Theresia Gschwandtner, Wolfgang Aigner, Katharina Kaiser, Silvia Miksch, and Andreas Seyfang. “CareCruiser: Exploring and Visualizing Plans, Events, and Effects Interactively”. In: *IEEE Pacific Visualization Symposium 2011, PacificVis 2011 - Proceedings*. 2011, pp. 43–50. ISBN: 978-1-61284-932-4. DOI: [10.1109/PACIFICVIS.2011.5742371](https://doi.org/10.1109/PACIFICVIS.2011.5742371).
- [93] George Hripcsak, David J. Albers, and Adler Perotte. “Exploiting Time in Electronic Health Record Correlations”. In: *Journal of the American Medical Informatics Association* 18.Supplement\_1 (Dec. 2011), pp. i109–i115. ISSN: 1527-974X. DOI: [10.1136/amiajnl-2011-000463](https://doi.org/10.1136/amiajnl-2011-000463).
- [94] Ryo Inoue. “A New Construction Method for Circle Cartograms”. In: *Cartography and Geographic Information Science* 38.2 (Jan. 2011), pp. 146–152. ISSN: 1523-0406, 1545-0465. DOI: [10.1559/15230406382146](https://doi.org/10.1559/15230406382146). (Visited on 02/23/2023).
- [95] R S Laramée. “How to Read a Visualization Research Paper: Extracting the Essentials”. In: *IEEE Computer Graphics and Applications* 31.3 (May 2011), pp. 78–82. ISSN: 0272-1716. DOI: [10.1109/MCG.2011.44](https://doi.org/10.1109/MCG.2011.44).
- [96] Sheri L. Lewis, Brian H. Feighner, Wayne A. Loschen, Richard A. Wojcik, Joseph F. Skora, Jacqueline S. Coberly, and David L. Blazes. “SAGES: A Suite of Freely-Available Software Tools for Electronic Disease Surveillance in Resource-Limited Settings”. In: *PLoS ONE* 6.5 (May 2011). Ed. by Abdisalan Mohamed Noor, e19750. ISSN: 1932-6203. DOI: [10.1371/journal.pone.0019750](https://doi.org/10.1371/journal.pone.0019750).
- [97] Ross Maciejewski, Philip Livengood, Stephen Rudolph, Timothy F. Collins, David S. Ebert, Robert T. Brigantic, Courtney D. Corley, George A. Muller, and Stephen W. Sanders. “A Pandemic Influenza Modeling and Visualization Tool”. In: *Journal of Visual Languages & Computing* 22.4 (Aug. 2011), pp. 268–278. ISSN: 1045926X. DOI: [10.1016/j.jvlc.2011.04.002](https://doi.org/10.1016/j.jvlc.2011.04.002).
- [98] *VAST Challenge 2011 MC1 - Characterization of an Epidemic Spread*. <https://old.datahub.io/dataset/vast-challenge-2011-mc1-characterization-of-an-epidemic-spread>. 2011.
- [99] H. Volzke, D. Alte, C. O. Schmidt, D. Radke, R. Lorbeer, N. Friedrich, N. Aumann, K. Lau, M. Piontek, G. Born, C. Havemann, T. Ittermann, S. Schipf, R. Haring, S. E. Baumeister, H. Wallaschofski, M. Nauck, S. Frick, A. Arnold, M. Junger, J. Mayerle, M. Kraft, M. M. Lerch, M. Dorr, T. Reffellmann, K. Empen, S. B. Felix, A. Obst, B. Koch, S. Glaser, R. Ewert, I. Fietze, T. Penzel, M. Doren, W. Rathmann, J. Haerting, M. Hannemann, J. Ropcke, U. Schminke, C. Jurgens, F. Tost, R. Rettig, J. A. Kors, S. Ungerer, K. Hegenscheid, J.-P. Kuhn, J. Kuhn, N. Hosten, R. Puls, J. Henke, O. Gloger, A. Teumer, G. Homuth, U. Volker, C. Schwahn, B. Holtfreter, I. Polzer, T. Kohlmann, H. J. Grabe, D. Rosskopf, H. K. Kroemer, T. Kocher, R. Biffar, U. John, and W. Hoffmann. “Cohort Profile: The Study of Health in Pomerania”. In: *International Journal of Epidemiology* 40.2 (Apr. 2011), pp. 294–307. ISSN: 0300-5771. DOI: [10.1093/ije/dyp394](https://doi.org/10.1093/ije/dyp394).

- [100] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. “LifeFlow: Visualizing an Overview of Event Sequences”. In: *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11*. Vancouver, BC, Canada: ACM Press, 2011, p. 1747. ISBN: 978-1-4503-0228-9. DOI: [10.1145/1978942.1979196](https://doi.org/10.1145/1978942.1979196). (Visited on 11/24/2020).
- [101] Krist Wongsuphasawat, John Alexis Guerra Gómez, Catherine Plaisant, Taowei David Wang, Meirav Taieb-Maimon, and Ben Shneiderman. “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: ACM Press, 2011, p. 1747. ISBN: 978-1-4503-0228-9. DOI: [10.1145/1978942.1979196](https://doi.org/10.1145/1978942.1979196).
- [102] Zhiyuan Zhang, Faisal Ahmed, Arunesh Mittal, IV Ramakrishnan, Rong Zhao, Asa Viccellio, and Klaus Mueller. “AnamneVis: A Framework for the Visualization of Patient History and Medical Diagnostics Chains”. In: *IEEE VAHC Workshop* January (2011), pp. 1–4.
- [103] Wladimir J. Alonso and Benjamin J.J. McCormick. “EPIPOI: A User-Friendly Analytical Tool for the Extraction and Visualization of Temporal Parameters from Epidemiological Time Series”. In: *BMC Public Health* 12.1 (Dec. 2012), p. 982. ISSN: 1471-2458. DOI: [10.1186/1471-2458-12-982](https://doi.org/10.1186/1471-2458-12-982).
- [104] Per Hans Gesteland, Yarden Livnat, Nathan Galli, Matthew H. Samore, and Adi V. Gundlapalli. “The EpiCanvas Infectious Disease Weather Map: An Interactive Visual Exploration of Temporal and Spatial Correlations”. In: *Journal of the American Medical Informatics Association* 19.6 (Nov. 2012), pp. 954–959. ISSN: 1067-5027. DOI: [10.1136/amiajnl-2011-000486](https://doi.org/10.1136/amiajnl-2011-000486).
- [105] Julian Heinrich and Daniel Weiskopf. “State of the Art of Parallel Coordinates”. In: *Eurographics 2013 - State of the Art Reports* (2012), 22 pages. ISSN: 1017-4656. DOI: [10.2312/CONF/EG2013/STARS/095-116](https://doi.org/10.2312/CONF/EG2013/STARS/095-116). (Visited on 03/15/2024).
- [106] Julie A. Jacko, ed. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. 3rd ed. Human Factors and Ergonomics. Boca Raton, FL: CRC Press, 2012. ISBN: 978-1-4398-2943-1.
- [107] Rohit Joshi and Peter Szolovits. “Prognostic Physiology: Modeling Patient Severity in Intensive Care Units Using Radial Domain Folding.” In: *AMIA ... Annual Symposium proceedings. AMIA Symposium 2012* (2012), pp. 1276–83. ISSN: 1942-597X.
- [108] Yarden Livnat, T. Rhyne, and Matthew Samore. “Epinome: A Visual-Analytics Workbench for Epidemiology Data”. In: *IEEE Computer Graphics and Applications* 32.2 (Mar. 2012), pp. 89–95. ISSN: 0272-1716. DOI: [10.1109/MCG.2012.31](https://doi.org/10.1109/MCG.2012.31).
- [109] Ketan K. Mane, Chris Bizon, Charles Schmitt, Phillips Owen, Bruce Burchett, Ricardo Pietrobon, and Kenneth Gersing. “VisualDecisionLinc: A Visual Analytics Approach for Comparative Effectiveness-Based Clinical Decision Support in Psychiatry”. In: *Journal of Biomedical Informatics* 45.1 (Feb. 2012), pp. 101–106. ISSN: 15320464. DOI: [10.1016/j.jbi.2011.09.003](https://doi.org/10.1016/j.jbi.2011.09.003).
- [110] Megan Monroe, Krist Wongsuphasawat, Catherine Plaisant, Ben Shneiderman, Jeff Millstein, and Sigfried Gold. “Exploring Point and Interval Event Patterns: Display Methods and Interactive Visual Query”. In: *HCIL Technical Report, Dept Computer Science, University of Maryland* May (2012), pp. 1–10.

- [111] Adam Perer and Jimeng Sun. “MatrixFlow: Temporal Network Visual Analytics to Track Symptom Evolution during Disease Progression.” In: *AMIA ... Annual Symposium proceedings. AMIA Symposium 2012* (2012), pp. 716–25. ISSN: 1942-597X.
- [112] H. Ribicic, J. Waser, R. Gurbat, B. Sadransky, and M. E. Groller. “Sketching Uncertainty into Simulations”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2255–2264. ISSN: 1077-2626. DOI: [10.1109/TVCG.2012.261](https://doi.org/10.1109/TVCG.2012.261). (Visited on 06/01/2023).
- [113] Awalin Sopan, Angela Song Ie Noh, Sohit Karol, Paul Rosenfeld, Ginnah Lee, and Ben Shneiderman. “Community Health Map: A Geospatial and Multivariate Data Visualization Tool for Public Health Datasets”. In: *Government Information Quarterly* 29.2 (2012), pp. 223–234. ISSN: 0740624X. DOI: [10.1016/j.giq.2011.10.002](https://doi.org/10.1016/j.giq.2011.10.002).
- [114] Brendan Stubbs, David C. Kale, and Amar Das. “Sim•TwentyFive: An Interactive Visualization System for Data-Driven Decision Support.” In: *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2012* (2012), pp. 891–900. ISSN: 1942597X.
- [115] K. Wongsuphasawat and D. Gotz. “Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2659–2668. ISSN: 1077-2626. DOI: [10.1109/TVCG.2012.225](https://doi.org/10.1109/TVCG.2012.225). (Visited on 11/24/2020).
- [116] Krist Wongsuphasawat and David Gotz. “Exploring Flow, Factors, and Outcomes of Temporal Event Sequences with the Outflow Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 18.12 (Dec. 2012), pp. 2659–2668. ISSN: 1077-2626. DOI: [10.1109/TVCG.2012.225](https://doi.org/10.1109/TVCG.2012.225).
- [117] Richard M. Bergenstal, Andrew J. Ahmann, Timothy Bailey, Roy W. Beck, Joan Bissen, Bruce Buckingham, Larry Deeb, Robert H. Dolin, Satish K. Garg, Robin Goland, Irl B. Hirsch, David C. Klonoff, Davida F. Kruger, Glenn Matfin, Roger S. Mazze, Beth A. Olson, Christopher Parkin, Anne Peters, Margaret A. Powers, Henry Rodriguez, Phil Southerland, Ellie S. Strock, William Tamborlane, and David M. Wesley. “Recommendations for Standardizing Glucose Reporting and Analysis to Optimize Clinical Decision Making in Diabetes: The Ambulatory Glucose Profile”. In: *Journal of Diabetes Science and Technology* 7.2 (Mar. 2013), pp. 562–578. ISSN: 1932-2968, 1932-2968. DOI: [10.1177/193229681300700234](https://doi.org/10.1177/193229681300700234). (Visited on 08/04/2023).
- [118] Matthew Brehmer and Tamara Munzner. “A Multi-Level Typology of Abstract Visualization Tasks”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2376–2385. ISSN: 1077-2626. DOI: [10.1109/TVCG.2013.124](https://doi.org/10.1109/TVCG.2013.124). (Visited on 07/14/2024).
- [119] Peter J. Brockwell and Richard A. Davis. *Time Series: Theory and Methods*. 2. ed., softcover repr. Springer Series in Statistics. New York, N.Y: Springer, 2013. ISBN: 978-1-4419-0319-8.
- [120] Hamish Cunningham, Valentin Tablan, Angus Roberts, and Kalina Bontcheva. “Getting More Out of Biomedical Documents with GATE’s Full Lifecycle Open Source Text Analytics”. In: *PLoS Computational Biology* 9.2 (Feb. 2013). Ed. by Andreas Prlic, e1002854. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.1002854](https://doi.org/10.1371/journal.pcbi.1002854).

- [121] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Moller. “A Systematic Review on the Practice of Evaluating Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (Dec. 2013), pp. 2818–2827. ISSN: 1077-2626. DOI: [10.1109/TVCG.2013.126](https://doi.org/10.1109/TVCG.2013.126).
- [122] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. “Temporal Event Sequence Simplification”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2227–2236. ISSN: 10772626. DOI: [10.1109/TVCG.2013.200](https://doi.org/10.1109/TVCG.2013.200).
- [123] Megan Monroe, Rongjian Lan, Hanseung Lee, Catherine Plaisant, and Ben Shneiderman. “Temporal Event Sequence Simplification”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), pp. 2227–2236. ISSN: 10772626. DOI: [10.1109/TVCG.2013.200](https://doi.org/10.1109/TVCG.2013.200).
- [124] Yair G. Rajwan, Pamela W. Barclay, Theresa Lee, I-Fong Sun, Catherine Passaretti, and Harold Lehmann. “Visualizing Central Line-Associated Blood Stream Infection (CLABSI) Outcome Data to Health Care Consumers and Practitioners for Decision Making – Evaluation Study”. In: *Online Journal of Public Health Informatics* 5.2 (June 2013), pp. 1–18. ISSN: 19472579. DOI: [10.5210/ojphi.v5i2.4364](https://doi.org/10.5210/ojphi.v5i2.4364).
- [125] Lilia L. Ramírez-Ramírez, Yulia R. Gel, Mary Thompson, Eileen de Villa, and Matt McPherson. “A New Surveillance and Spatio-Temporal Visualization Tool SIMID: SIMulation of Infectious Diseases Using Random Networks and GIS”. In: *Computer Methods and Programs in Biomedicine* 110.3 (June 2013), pp. 455–470. ISSN: 01692607. DOI: [10.1016/j.cmpb.2013.01.007](https://doi.org/10.1016/j.cmpb.2013.01.007).
- [126] Alexander Rind, Taowei David Wang, Wolfgang Aigner, Silvia Miksch, Krist Wongsuphasawat, Catherine Plaisant, and Ben Shneiderman. “Interactive Information Visualization to Explore and Query Electronic Health Records”. In: *Foundations and Trends® in Human-Computer Interaction* 5.3 (Feb. 2013), pp. 207–298. ISSN: 1551-3955, 1551-3963. DOI: [10.1561/11000000039](https://doi.org/10.1561/11000000039). (Visited on 11/24/2020).
- [127] Zhiyuan Zhang, Bing Wang, Faisal Ahmed, I. V. Ramakrishnan, Rong Zhao, Asa Viccellio, and Klaus Mueller. “The Five Ws for Information Visualization with Application to Healthcare Informatics”. In: *IEEE Transactions on Visualization and Computer Graphics* 19.11 (Nov. 2013), pp. 1895–1910. DOI: [10.1109/TVCG.2013.89](https://doi.org/10.1109/TVCG.2013.89).
- [128] David Borland, Vivian L West, and W Ed Hammond. “Multivariate Visualization of System-Wide National Health Service Data Using Radial Coordinates”. In: *Proceedings of the 2014 Workshop on Visual Analytics in Healthcare* (2014), pp. 19–24.
- [129] Lauren N. Carroll, Alan P. Au, Landon Todd Detwiler, Tsung-chieh Fu, Ian S. Painter, and Neil F. Abernethy. “Visualization and Analytics Tools for Infectious Disease Epidemiology: A Systematic Review”. In: *Journal of Biomedical Informatics* 51.1 (Oct. 2014), pp. 287–298. ISSN: 15320464. DOI: [10.1016/j.jbi.2014.04.006](https://doi.org/10.1016/j.jbi.2014.04.006).
- [130] C.C. Freifeld, K.D. Mandl, B.Y. Reis, and J.S. Brownstein. “HealthMap : Global Infectious Disease Monitoring Through”. In: *Journal of the American Medical Informatics Association* 15.2 (2014), pp. 150–157. ISSN: 10675027. DOI: [10.1197/jamia.M2544.Introduction](https://doi.org/10.1197/jamia.M2544.Introduction).

- [131] Jorge A Gálvez, Luis Ahumada, Allan F Simpao, Elaina E Lin, Christopher P Bonafide, Dhruv Choudhry, William R England, Abbas F Jawad, David Friedman, Debora A Sesok-Pizzini, and Mohamed A Rehman. “Visual Analytical Tool for Evaluation of 10-Year Perioperative Transfusion Practice at a Children’s Hospital”. In: *Journal of the American Medical Informatics Association* 21.3 (May 2014), pp. 529–534. ISSN: 1067-5027. DOI: [10.1136/amiajnl-2013-002241](https://doi.org/10.1136/amiajnl-2013-002241).
- [132] David Gotz and Harry Stavropoulos. “DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 1783–1792. ISSN: 10772626. DOI: [10.1109/TVCG.2014.2346682](https://doi.org/10.1109/TVCG.2014.2346682).
- [133] Jaemin Jo, Jaeseok Huh, Jonghun Park, Bohyoung Kim, and Jinwook Seo. “LiveGantt: Interactively Visualizing a Large Manufacturing Schedule”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (Dec. 2014), pp. 2329–2338. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2014.2346454](https://doi.org/10.1109/TVCG.2014.2346454). (Visited on 06/05/2023).
- [134] Rishikesan Kamaleswaran, James Edward Pugh, Anirudh Thommandram, Andrew James, and Carolyn McGregor. “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.
- [135] Artem Konev, Jürgen Waser, Bernhard Sadransky, Daniel Cornel, Rui A.P. Perdigão, Zsolt Horváth, and M. Eduard Gröller. “Run Watchers: Automatic Simulation-Based Decision Support in Flood Management”. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (Dec. 2014), pp. 1873–1882. ISSN: 1941-0506. DOI: [10.1109/TVCG.2014.2346930](https://doi.org/10.1109/TVCG.2014.2346930).
- [136] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. “An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences”. In: *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014), pp. 1–6.
- [137] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. “An Evaluation of Visual Analytics Approaches to Comparing Cohorts of Event Sequences”. In: *Proc. of IEEE VIS 2014 Workshop on Visualization of Electronic Health Records* (2014), pp. 1–6.
- [138] Tamara Munzner. *Visualization Analysis and Design*. 1st ed. New York: A K Peters/CRC Press, Dec. 2014. ISBN: 978-0-429-08890-2. DOI: [10.1201/b17511](https://doi.org/10.1201/b17511). (Visited on 04/12/2023).
- [139] W. Owen Pickrell, Arron S. Lacey, Rhys H. Thomas, Ronan A. Lyons, Phil E.M. Smith, and Mark I. Rees. “Trends in the First Antiepileptic Drug Prescribed for Epilepsy between 2000 and 2010”. In: *Seizure* 23.1 (Jan. 2014), pp. 77–80. ISSN: 10591311. DOI: [10.1016/j.seizure.2013.09.007](https://doi.org/10.1016/j.seizure.2013.09.007). (Visited on 07/14/2024).
- [140] Allan F. Simpao, Luis M. Ahumada, Bimal R. Desai, Christopher P. Bonafide, J. A. Galvez, Mohamed A. Rehman, Abbas F. Jawad, Krisha L. Palma, and Eric D. Shelov. “Optimization of Drug-Drug Interaction Alert Rules in a Pediatric Hospital’s Electronic Health Record System Using a Visual Analytics Dashboard”. In: *Journal of the American Medical Informatics Association* 22.2 (Oct. 2014), pp. 361–369. ISSN: 1067-5027. DOI: [10.1136/amiajnl-2013-002538](https://doi.org/10.1136/amiajnl-2013-002538).



- [141] Rajeev Agrawal, Anirudh Kadadi, Xiangfeng Dai, and Frederic Andres. “Challenges and Opportunities with Big Data Visualization”. In: *Proceedings of the 7th International Conference on Management of Computational and Collective intelligence in Digital EcoSystems*. Caraguatatuba Brazil: ACM, Oct. 2015, pp. 169–173. ISBN: 978-1-4503-3480-8. DOI: [10.1145/2857218.2857256](https://doi.org/10.1145/2857218.2857256). (Visited on 07/10/2022).
- [142] Md. Jawaherul Alam, Stephen G. Kobourov, and Sankar Veeramoni. “Quantitative Measures for Cartogram Generation Techniques”. In: *Computer Graphics Forum* 34.3 (June 2015), pp. 351–360. ISSN: 01677055. DOI: [10.1111/cgf.12647](https://doi.org/10.1111/cgf.12647). (Visited on 02/23/2023).
- [143] Jurgen Bernard, David Sessler, Thorsten May, Thorsten Schlomm, Dirk Pehrke, and Jorn Kohlhammer. “A Visual-Interactive System for Prostate Cancer Cohort Analysis”. In: *IEEE Computer Graphics and Applications* 35.3 (May 2015), pp. 44–55. ISSN: 0272-1716. DOI: [10.1109/MCG.2015.49](https://doi.org/10.1109/MCG.2015.49). (Visited on 11/24/2020).
- [144] Jürgen Bernard, David Sessler, Andreas Bannach, Thorsten May, and Jörn Kohlhammer. “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*. Chicago, Illinois: ACM Press, 2015, pp. 1–8. ISBN: 978-1-4503-3671-0. DOI: [10.1145/2836034.2836035](https://doi.org/10.1145/2836034.2836035). (Visited on 11/24/2020).
- [145] Jürgen Bernard, David Sessler, Andreas Bannach, Thorsten May, and Jörn Kohlhammer. “A Visual Active Learning System for the Assessment of Patient Well-Being in Prostate Cancer Research”. In: vol. 25-October. *Proceedings of the 2015 Workshop on Visual Analytics in Healthcare - VAHC '15*. ACM Press, 2015, pp. 1–8. ISBN: 978-1-4503-3671-0. DOI: [10.1145/2836034.2836035](https://doi.org/10.1145/2836034.2836035).
- [146] R. G. Cano, K. Buchin, T. Castermans, A. Pieterse, W. Sonke, and B. Speckmann. “Mosaic Drawings and Cartograms”. In: *Computer Graphics Forum* 34.3 (June 2015), pp. 361–370. ISSN: 01677055. DOI: [10.1111/cgf.12648](https://doi.org/10.1111/cgf.12648). (Visited on 01/15/2022).
- [147] Cody Dunne, Michael Muller, Nicola Perra, and Mauro Martino. “VoroGraph: Visualization Tools for Epidemic Analysis”. In: *Conference on Human Factors in Computing Systems - Proceedings* 18 (2015), pp. 255–258. DOI: [10.1145/2702613.2725459](https://doi.org/10.1145/2702613.2725459).
- [148] P Federico, J Unger, L Sacchi, D Klimov, and S Miksch. “Gnaeus : Utilizing Clinical Guidelines for Knowledge-Assisted Visualisation of EHR Cohorts”. In: *EuroVis Workshop on Visual Analytics* (2015). DOI: [10.2312/eurova.20151108](https://doi.org/10.2312/eurova.20151108).
- [149] The Apache Software Foundation. *Apache Lucene*. <https://lucene.apache.org/>. 2015.
- [150] Alistair Johnson, Tom Pollard, and Roger Mark. *MIMIC-III Clinical Database*. 2015. DOI: [10.13026/C2XW26](https://doi.org/10.13026/C2XW26). (Visited on 07/17/2022).
- [151] Paul Klemm, Kai Lawonn, Sylvia Glaber, Uli Niemann, Katrin Hegenscheid, Henry Volzke, and Bernhard Preim. “3D Regression Heat Map Analysis of Population Study Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2015), pp. 81–90. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2015.2468291](https://doi.org/10.1109/TVCG.2015.2468291). (Visited on 08/07/2022).

- [152] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. “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*. Vol. 2015-Janua. New York, New York, USA: ACM Press, 2015, pp. 38–49. ISBN: 978-1-4503-3306-1. DOI: [10.1145/2678025.2701407](https://doi.org/10.1145/2678025.2701407).
- [153] A. Névél, P. Zweigenbaum, and Section Editors for the IMIA Yearbook Section on Clinical Natural Language Processing. “Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare”. In: *Yearbook of Medical Informatics* 24.01 (Aug. 2015), pp. 194–198. ISSN: 0943-4747, 2364-0502. DOI: [10.15265/IY-2015-035](https://doi.org/10.15265/IY-2015-035). (Visited on 07/14/2024).
- [154] Adam Perer, Fei Wang, and Jianying Hu. “Mining and Exploring Care Pathways from Electronic Medical Records with Visual Analytics”. In: *Journal of Biomedical Informatics* 56 (Aug. 2015), pp. 369–378. ISSN: 15320464. DOI: [10.1016/j.jbi.2015.06.020](https://doi.org/10.1016/j.jbi.2015.06.020). (Visited on 08/28/2022).
- [155] Sunghwan Suh and Jae Hyeon Kim. “Glycemic Variability: How Do We Measure It and Why Is It Important?” In: *Diabetes & Metabolism Journal* 39.4 (2015), p. 273. ISSN: 2233-6079, 2233-6087. DOI: [10.4093/dmj.2015.39.4.273](https://doi.org/10.4093/dmj.2015.39.4.273). (Visited on 10/06/2023).
- [156] IHME University of Washington. *GBD Compare*. <http://vizhub.healthdata.org/gbd-compare>. 2015.
- [157] Vivian L West, David Borland, and W Ed Hammond. “Innovative Information Visualization of Electronic Health Record Data: A Systematic Review”. In: *Journal of the American Medical Informatics Association* 22.2 (Mar. 2015), pp. 330–339. ISSN: 1527974X. DOI: [10.1136/amiajnl-2014-002955](https://doi.org/10.1136/amiajnl-2014-002955).
- [158] Max Bearak and Lazaro Gamio. “Everything You Ever Wanted to Know about the U.S. Foreign Assistance Budget”. In: *Washington Post* (Oct. 2016). <https://www.washingtonpost.com/graphics/world/which-countries-get-the-most-foreign-aid/>. (Visited on 01/16/2022).
- [159] Nan Cao and Weiwei Cui. *Introduction to Text Visualization*. Paris: Atlantis Press, 2016. ISBN: 978-94-6239-185-7 978-94-6239-186-4. DOI: [10.2991/978-94-6239-186-4](https://doi.org/10.2991/978-94-6239-186-4). (Visited on 09/22/2024).
- [160] Flávio Dusse, Paulo Simões Júnior, Antonia Tamires Alves, Renato Novais, Vaninha Vieira, and Manoel Mendonça. “Information Visualization for Emergency Management: A Systematic Mapping Study”. In: *Expert Systems with Applications* 45 (Mar. 2016), pp. 424–437. ISSN: 0957-4174. DOI: [10.1016/j.eswa.2015.10.007](https://doi.org/10.1016/j.eswa.2015.10.007). (Visited on 06/01/2023).
- [161] R. S. Evans. “Electronic Health Records: Then, Now, and in the Future”. In: *Yearbook of Medical Informatics* 25.S 01 (Aug. 2016), S48–S61. ISSN: 0943-4747, 2364-0502. DOI: [10.15265/IYS-2016-s006](https://doi.org/10.15265/IYS-2016-s006). (Visited on 11/24/2020).
- [162] Lazaro Gamio. “Election Maps Are Telling You Big Lies about Small Things”. In: *Washington Post* (Nov. 2016). <https://www.washingtonpost.com/graphics/politics/2016-election/how-election-maps-lie/>. (Visited on 01/16/2022).
- [163] Michael Glueck, Peter Hamilton, Fanny Chevalier, Simon Breslav, Azam Khan, Daniel Wigdor, and Michael Brudno. “PhenoBlocks: Phenotype Comparison Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 101–110. ISSN: 1077-2626. DOI: [10.1109/TVCG.2015.2467733](https://doi.org/10.1109/TVCG.2015.2467733). (Visited on 11/24/2020).

- [164] David Gotz and David Borland. “Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization”. In: *IEEE Computer Graphics and Applications* 36.3 (May 2016), pp. 90–96. ISSN: 0272-1716. DOI: [10.1109/MCG.2016.59](https://doi.org/10.1109/MCG.2016.59).
- [165] Trevor Hogan, Uta Hinrichs, and Eva Hornecker. “The Elicitation Interview Technique: Capturing People’s Experiences of Data Representations”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.12 (Dec. 2016), pp. 2579–2593. ISSN: 1077-2626. DOI: [10.1109/TVCG.2015.2511718](https://doi.org/10.1109/TVCG.2015.2511718). (Visited on 05/14/2021).
- [166] Shenhui Jiang, Shiao-fen Fang, Sam Bloomquist, Jeremy Keiper, Mathew Palakal, Yuni Xia, and Shaun Grannis. “Healthcare Data Visualization: Geospatial and Temporal Integration”. In: *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. Vol. 2. SCITEPRESS - Science and Technology Publications, 2016, pp. 212–219. ISBN: 978-989-758-175-5. DOI: [10.5220/0005714002120219](https://doi.org/10.5220/0005714002120219).
- [167] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. “MIMIC-III, a Freely Accessible Critical Care Database”. In: *Scientific Data* 3.1 (Dec. 2016), p. 160035. ISSN: 2052-4463. DOI: [10.1038/sdata.2016.35](https://doi.org/10.1038/sdata.2016.35). (Visited on 05/14/2021).
- [168] Rishikesan Kamaleswaran, Christopher Collins, Andrew James, and Carolyn McGregor. “PhysioEx: Visual Analysis of Physiological Event Streams”. In: *Computer Graphics Forum* 35.3 (2016), pp. 331–340. ISSN: 14678659. DOI: [10.1111/cgf.12909](https://doi.org/10.1111/cgf.12909).
- [169] Rishikesan Kamaleswaran, Christopher Collins, Andrew James, and Carolyn McGregor. “PhysioEx: Visual Analysis of Physiological Event Streams”. In: *Computer Graphics Forum* 35.3 (2016), pp. 331–340. ISSN: 14678659. DOI: [10.1111/cgf.12909](https://doi.org/10.1111/cgf.12909).
- [170] Bum Chul Kwon, Janu Verma, and Adam Perer. “Peekquence: Visual Analytics for Event Sequence Data”. In: *ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics (IDEA’16)* 1 (Aug. 2016).
- [171] Mona Hosseinkhani Looarak, Charles Perin, Noreen Kamal, Michael Hill, and Sheelagh Carpendale. “TimeSpan: Using Visualization to Explore Temporal Multi-dimensional Data of Stroke Patients”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), pp. 409–418. ISSN: 10772626. DOI: [10.1109/TVCG.2015.2467325](https://doi.org/10.1109/TVCG.2015.2467325).
- [172] Masood Masoodian, Saturnino Luz, and David Kavenga. “Nu-View: A Visualization System for Collaborative Co-located Analysis of Geospatial Disease Data”. In: *ACM International Conference Proceeding Series* 01-05-Febr (2016). DOI: [10.1145/2843043.2843374](https://doi.org/10.1145/2843043.2843374).
- [173] Matthew Louis Mauriello, Ben Shneiderman, Fan Du, Sana Malik, and Catherine Plaisant. “Simplifying Overviews of Temporal Event Sequences”. In: *Conference on Human Factors in Computing Systems - Proceedings* 07-12-May (2016), pp. 2217–2224. DOI: [10.1145/2851581.2892440](https://doi.org/10.1145/2851581.2892440).
- [174] Greg Miller. “Election Maps Can Be Misleading—Here’s a Solution”. In: *National Geographic* (Oct. 2016). <https://www.nationalgeographic.com/culture/article/improved-election-map-cartograms>. (Visited on 01/16/2022).



- [175] MIT Critical Data. *Secondary Analysis of Electronic Health Records*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-43740-8 978-3-319-43742-2. DOI: [10.1007/978-3-319-43742-2](https://doi.org/10.1007/978-3-319-43742-2). (Visited on 08/07/2022).
- [176] Sabrina Nusrat and Stephen Kobourov. “The State of the Art in Cartograms”. In: *Computer Graphics Forum* 35.3 (June 2016), pp. 619–642. ISSN: 01677055. DOI: [10.1111/cgf.12932](https://doi.org/10.1111/cgf.12932). (Visited on 12/05/2021).
- [177] Oluwakemi Ola and Kamran Sedig. “Beyond Simple Charts: Design of Visualizations for Big Health Data”. In: *Online Journal of Public Health Informatics* 8.3 (Dec. 2016). ISSN: 1947-2579. DOI: [10.5210/ojphi.v8i3.7100](https://doi.org/10.5210/ojphi.v8i3.7100). (Visited on 11/24/2020).
- [178] Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. “Data Visualization Tools for Investigating Health Services Utilization Among Cancer Patients”. In: *Oncology Informatics*. Elsevier Inc., 2016, pp. 207–229. ISBN: 978-0-12-802115-6. DOI: [10.1016/b978-0-12-802115-6.00011-2](https://doi.org/10.1016/b978-0-12-802115-6.00011-2).
- [179] 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)*. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504&from=EN>. 2016.
- [180] Martin Stabe. “The Search for a Better US Election Map”. In: *Financial Times* (Nov. 2016). <https://www.ft.com/content/3685bf9e-a4cc-11e6-8b69-02899e8bd9d1>. (Visited on 01/16/2022).
- [181] Adam Taylor. “What a Real ‘Brexit Britain’ Would Look Like”. In: *Washington Post* (June 2016). <https://www.washingtonpost.com/news/worldviews/wp/2016/06/29/what-a-real-brexit-britain-would-look-like/>. (Visited on 01/16/2022).
- [182] Xu Meng Wang, Tian Ye Zhang, Yu Xin Ma, Jing Xia, and Wei Chen. “A Survey of Visual Analytic Pipelines”. In: *Journal of Computer Science and Technology* 31.4 (2016), pp. 787–804. ISSN: 10009000. DOI: [10.1007/s11390-016-1663-1](https://doi.org/10.1007/s11390-016-1663-1).
- [183] Wenchao Wu, Jiayi Xu, Haipeng Zeng, Yixian Zheng, Huamin Qu, Bing Ni, Mingxuan Yuan, and Lionel M. Ni. “TelCoVis: Visual Exploration of Co-occurrence in Urban Human Mobility Based on Telco Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 22.1 (Jan. 2016), pp. 935–944. ISSN: 1077-2626. DOI: [10.1109/TVCG.2015.2467194](https://doi.org/10.1109/TVCG.2015.2467194). (Visited on 10/27/2023).
- [184] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. “Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. Barcelona, Spain: IEEE, Dec. 2016, pp. 1317–1322. ISBN: 978-1-5090-5473-2. DOI: [10.1109/ICDM.2016.0179](https://doi.org/10.1109/ICDM.2016.0179). (Visited on 05/12/2023).
- [185] Yan Zhu, Zachary Zimmerman, Nader Shakibay Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. “Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins”. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. Barcelona, Spain: IEEE, Dec. 2016, pp. 739–748. ISBN: 978-1-5090-5473-2. DOI: [10.1109/ICDM.2016.0085](https://doi.org/10.1109/ICDM.2016.0085). (Visited on 05/12/2023).

- [186] Humberto S. Garcia Caballero, Alberto Corvo, Prabhakar M. Dixit, and Michel A. Westenberg. “Visual Analytics for Evaluating Clinical Pathways”. In: *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*. IEEE, Oct. 2017, pp. 39–46. ISBN: 978-1-5386-3187-4. DOI: [10.1109/VAHC.2017.8387499](https://doi.org/10.1109/VAHC.2017.8387499).
- [187] Martin R. Cowie, Juuso I. Blomster, Lesley H. Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, Alexander Michel, Seleen Ong, Jill P. Pell, Mary Ross Southworth, Wendy Gattis Stough, Martin Thoenes, Faiez Zannad, and Andrew Zalewski. “Electronic Health Records to Facilitate Clinical Research”. In: *Clinical Research in Cardiology* 106.1 (2017), pp. 1–9. ISSN: 18610692. DOI: [10.1007/s00392-016-1025-6](https://doi.org/10.1007/s00392-016-1025-6).
- [188] Pedro Cruz. “Adapted Dorling Cartogram on Wage Inequality in Portugal”. In: *2017 IEEE VIS Arts Program (VISAP)*. Phoenix, AZ: IEEE, Oct. 2017, pp. 1–5. ISBN: 978-1-5386-3490-5. DOI: [10.1109/VISAP.2017.8282365](https://doi.org/10.1109/VISAP.2017.8282365). (Visited on 12/08/2021).
- [189] Filip Dabek, Elizabeth Jimenez, and Jesus J. Caban. “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. ISBN: 978-1-5386-3187-4. DOI: [10.1109/VAHC.2017.8387501](https://doi.org/10.1109/VAHC.2017.8387501).
- [190] Fan Du, Ben Shneiderman, Catherine Plaisant, Sana Malik, and Adam Perer. “Coping with Volume and Variety in Temporal Event Sequences: Strategies for Sharpening Analytic Focus”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.6 (June 2017), pp. 1636–1649. ISSN: 1077-2626. DOI: [10.1109/TVCG.2016.2539960](https://doi.org/10.1109/TVCG.2016.2539960).
- [191] Dheeru Dua and Casey Graff. *{UCI} Machine Learning Repository*. <http://archive.ics.uci.edu/ml/>. 2017.
- [192] Jaillah Mae Gesulga, Almarie Berjame, Kristelle Sheen Moquiala, and Adrian Galido. “Barriers to Electronic Health Record System Implementation and Information Systems Resources: A Structured Review”. In: *Procedia Computer Science* 124 (2017), pp. 544–551. ISSN: 18770509. DOI: [10.1016/j.procs.2017.12.188](https://doi.org/10.1016/j.procs.2017.12.188). (Visited on 11/24/2020).
- [193] Michael Glueck, Alina Gvozdk, Fanny Chevalier, Azam Khan, Michael Brudno, and Daniel Wigdor. “PhenoStacks: Cross-sectional Cohort Phenotype Comparison Visualizations”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), pp. 191–200. ISSN: 10772626. DOI: [10.1109/TVCG.2016.2598469](https://doi.org/10.1109/TVCG.2016.2598469).
- [194] Connor C. Gramazio, David H. Laidlaw, and Karen B. Schloss. “Colorgorical: Creating Discriminable and Preferable Color Palettes for Information Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.1 (Jan. 2017), pp. 521–530. ISSN: 1941-0506. DOI: [10.1109/TVCG.2016.2598918](https://doi.org/10.1109/TVCG.2016.2598918). (Visited on 08/10/2024).
- [195] Tsipi Heart, Ofir Ben-Assuli, and Itamar Shabtai. “A Review of PHR, EMR and EHR Integration: A More Personalized Healthcare and Public Health Policy”. In: *Health Policy and Technology* 6.1 (Mar. 2017), pp. 20–25. ISSN: 22118837. DOI: [10.1016/j.hlpt.2016.08.002](https://doi.org/10.1016/j.hlpt.2016.08.002).

- [196] Petra Isenberg, Florian Heimerl, Steffen Koch, Tobias Isenberg, Panpan Xu, Charles D. Stolper, Michael Sedlmair, Jian Chen, Torsten Moller, and John Stasko. “Vispubdata.Org: A Metadata Collection about IEEE Visualization (VIS) Publications”. In: *IEEE Transactions on Visualization and Computer Graphics* 23.9 (Sept. 2017), pp. 2199–2206. ISSN: 1077-2626. DOI: [10.1109/TVCG.2016.2615308](https://doi.org/10.1109/TVCG.2016.2615308).
- [197] Liam McNabb and Robert S. Laramee. “Survey of Surveys (SoS) - Mapping The Landscape of Survey Papers in Information Visualization”. In: *Computer Graphics Forum* 36.3 (June 2017), pp. 589–617. ISSN: 0167-7055. DOI: [10.1111/cgf.13212](https://doi.org/10.1111/cgf.13212).
- [198] Liam McNabb and Robert S. Laramee. “Survey of Surveys (SoS) - Mapping the Landscape of Survey Papers in Information Visualization”. In: vol. 36. Computer Graphics Forum. The Eurographics Association & John Wiley & Sons, Ltd., June 2017, pp. 589–617. DOI: [10.1111/cgf.13212](https://doi.org/10.1111/cgf.13212).
- [199] World Health Organization and Pan American Health Organization. “Handbook for Electronic Health Records Implementation”. In: June (2017). [http://www.paho.org/ict4health/images/docs/DRAFT-Handbook\\_EHR\\_Implementation.pdf](http://www.paho.org/ict4health/images/docs/DRAFT-Handbook_EHR_Implementation.pdf), p. 75.
- [200] Alexander Rind, Paolo Federico, Theresia Gschwandtner, Wolfgang Aigner, Jakob Doppler, and Markus Wagner. “Visual Analytics of Electronic Health Records with a Focus on Time”. In: *New Perspectives in Medical Records*. Ed. by Fabio Capello, Giovanni Rinaldi, and Giovanna Gatti. 2017. Chap. 5, pp. 65–77. ISBN: 978-3-319-28659-4. DOI: [10.1007/978-3-319-28661-7](https://doi.org/10.1007/978-3-319-28661-7).
- [201] Chao Tong, Liam McNabb, Robert S. Laramee, Jane Lyons, Angharad Walters, Damon Berridge, and Daniel Thayer. “Time-Oriented Cartographic Treemaps for Visualization of Public Healthcare Data”. In: *Computer Graphics and Visual Computing (CGVC)*. 2017, 18 pages. ISBN: 9783038680505. DOI: [10.2312/CGVC.20171273](https://doi.org/10.2312/CGVC.20171273). (Visited on 01/29/2025).
- [202] Chao Tong, Richard Roberts, Robert S. Laramee, Damon Berridge, and Daniel Thayer. “Cartographic Treemaps for Visualization of Public Healthcare Data”. In: *Computer Graphics and Visual Computing (CGVC)*. The Eurographics Association, 2017, 14 pages. ISBN: 9783038680505. DOI: [10.2312/CGVC.20171276](https://doi.org/10.2312/CGVC.20171276). (Visited on 01/29/2025).
- [203] Guy J. Abel. “Estimates of Global Bilateral Migration Flows by Gender between 1960 and 2015<sup>1</sup>”. In: *International Migration Review* 52.3 (Sept. 2018), pp. 809–852. ISSN: 0197-9183, 1747-7379. DOI: [10.1111/imre.12327](https://doi.org/10.1111/imre.12327). (Visited on 07/10/2022).
- [204] Felwa Abukhodair, Khalid Khashoggi, Tim O’Connell, and Chris Shaw. “Rad-Stream: An Interactive Visual Display of Radiology Workflow for Delay Detection in the Clinical Imaging Process”. In: *2017 IEEE Workshop on Visual Analytics in Healthcare, VAHC 2017* (2018), pp. 69–76. DOI: [10.1109/VAHC.2017.8387543](https://doi.org/10.1109/VAHC.2017.8387543).
- [205] Anthony Breitzman. “Using Cartograms to Visualize Population Normalized Big-Data Sets”. In: *2018 IEEE International Conference on Big Data (Big Data)*. Seattle, WA, USA: IEEE, Dec. 2018, pp. 3575–3580. ISBN: 978-1-5386-5035-6. DOI: [10.1109/BigData.2018.8622217](https://doi.org/10.1109/BigData.2018.8622217). (Visited on 03/03/2022).

- [206] Yuanzhe Chen, Panpan Xu, and Liu Ren. “Sequence Synopsis: Optimize Visual Summary of Temporal Event Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), pp. 45–55. ISSN: 10772626. DOI: [10.1109/TVCG.2017.2745083](https://doi.org/10.1109/TVCG.2017.2745083).
- [207] VA Community. *VAST Challenge 2018*. <https://data.pnnl.gov/group/nodes/dataset/13217>. 2018.
- [208] Marta Galluzzi, Duccio Rocchini, Roberto Canullo, Ronald E. McRoberts, and Gherardo Chirici. “Mapping Uncertainty of ICP-Forest Biodiversity Data: From Standard Treatment of Diffusion to Density-Equalizing Cartograms”. In: *Ecological Informatics* 48 (Nov. 2018), pp. 281–289. ISSN: 15749541. DOI: [10.1016/j.ecoinf.2018.06.005](https://doi.org/10.1016/j.ecoinf.2018.06.005). (Visited on 03/03/2022).
- [209] Michael Glueck, Mahdi Pakdaman Naeni, Finale Doshi-Velez, Fanny Chevalier, Azam Khan, Daniel Wigdor, and Michael Brudno. “PhenoLines: Phenotype Comparison Visualizations for Disease Subtyping via Topic Models”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (Jan. 2018), pp. 371–381. ISSN: 1077-2626. DOI: [10.1109/TVCG.2017.2745118](https://doi.org/10.1109/TVCG.2017.2745118). (Visited on 11/24/2020).
- [210] Shunan Guo, Ke Xu, Rongwen Zhao, David Gotz, Hongyuan Zha, and Nan Cao. “EventThread: Visual Summarization and Stage Analysis of Event Sequence Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (Jan. 2018), pp. 56–65. ISSN: 1077-2626. DOI: [10.1109/TVCG.2017.2745320](https://doi.org/10.1109/TVCG.2017.2745320).
- [211] Richard Harris, Martin Charlton, and Chris Brunsdon. “Mapping the Changing Residential Geography of White British Secondary School Children in England Using Visually Balanced Cartograms and Hexograms”. In: *Journal of Maps* 14.1 (Jan. 2018), pp. 65–72. ISSN: 1744-5647. DOI: [10.1080/17445647.2018.1478753](https://doi.org/10.1080/17445647.2018.1478753). (Visited on 03/03/2022).
- [212] Arron S Lacey, William Owen Pickrell, Rhys H Thomas, Mike P Kerr, Cathy P White, and Mark I Rees. “Educational Attainment of Children Born to Mothers with Epilepsy”. In: *Journal of Neurology, Neurosurgery & Psychiatry* 89.7 (July 2018), pp. 736–740. ISSN: 0022-3050, 1468-330X. DOI: [10.1136/jnnp-2017-317515](https://doi.org/10.1136/jnnp-2017-317515). (Visited on 07/19/2024).
- [213] Liam McNabb, Robert S Laramée, and Richard Fry. “Dynamic Choropleth Maps – Using Amalgamation to Increase Area Perceivability”. In: *2018 22nd International Conference Information Visualisation (IV)*. Fisciano, Italy: IEEE, July 2018, pp. 284–293. ISBN: 978-1-5386-7202-0. DOI: [10.1109/iV.2018.00056](https://doi.org/10.1109/iV.2018.00056). (Visited on 07/10/2022).
- [214] Medicines and Healthcare products Regulatory Agency. *New Measures to Avoid Valproate Exposure in Pregnancy*. <https://www.gov.uk/government/news/new-measures-to-avoid-valproate-exposure-in-pregnancy>. 2018. (Visited on 07/14/2024).
- [215] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenbourn, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. “Scalable and Accurate Deep Learning with Electronic Health Records”. In: *npj Digital*

*Medicine* 1.1 (Dec. 2018), p. 18. ISSN: 2398-6352. DOI: [10.1038/s41746-018-0029-1](https://doi.org/10.1038/s41746-018-0029-1).

- [216] Michael Sandberg. *Cartogram: House Election Results: Democrats Take Control (The New York Times)*. <https://datavizblog.com/2018/11/14/cartogram-house-election-results-democrats-take-control-the-new-york-times/>. Nov. 2018. (Visited on 01/16/2022).
- [217] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. “Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis”. In: *IEEE journal of biomedical and health informatics* 22.5 (Sept. 2018), pp. 1589–1604. ISSN: 2168-2208. DOI: [10.1109/JBHI.2017.2767063](https://doi.org/10.1109/JBHI.2017.2767063).
- [218] Chao Tong, Liam McNabb, and Robert S. Laramée. “Cartograms with Topological Features”. In: *Computer Graphics and Visual Computing (CGVC)*. The Eurographics Association, 2018, 8 pages. ISBN: 9783038680710. DOI: [10.2312/CGVC.20181217](https://doi.org/10.2312/CGVC.20181217). (Visited on 01/29/2025).
- [219] Gaurav Trivedi, Phuong Pham, Wendy W Chapman, Rebecca Hwa, Janyce Wiebe, and Harry Hochheiser. “NLPreViz: An Interactive Tool for Natural Language Processing on Clinical Text”. In: *Journal of the American Medical Informatics Association* 25.1 (Jan. 2018), pp. 81–87. DOI: [10.1093/jamia/ocx070](https://doi.org/10.1093/jamia/ocx070).
- [220] Willem G van Panhuis, Anne Cross, and Donald S Burke. “Project Tycho 2.0: A Repository to Improve the Integration and Reuse of Data for Global Population Health”. In: *Journal of the American Medical Informatics Association* 25.12 (Dec. 2018), pp. 1608–1617. ISSN: 1067-5027. DOI: [10.1093/jamia/ocy123](https://doi.org/10.1093/jamia/ocy123).
- [221] Yan Zhu, Chin-Chia Michael Yeh, Zachary Zimmerman, Kaveh Kamgar, and Eamonn Keogh. “Matrix Profile XI: SCRIMP++: Time Series Motif Discovery at Interactive Speeds”. In: *2018 IEEE International Conference on Data Mining (ICDM)*. Singapore: IEEE, Nov. 2018, pp. 837–846. ISBN: 978-1-5386-9159-5. DOI: [10.1109/ICDM.2018.00099](https://doi.org/10.1109/ICDM.2018.00099). (Visited on 05/12/2023).
- [222] S. Alemzadeh, U. Niemann, T. Ittermann, H. Völzke, D. Schneider, M. Spiliopoulou, K. Bühler, and B. Preim. “Visual Analysis of Missing Values in Longitudinal Cohort Study Data”. In: *Computer Graphics Forum* 39.1 (May 2019), pp. 63–75. ISSN: 0167-7055, 1467-8659. DOI: [10.1111/cgf.13662](https://doi.org/10.1111/cgf.13662). (Visited on 08/07/2022).
- [223] Mohammed Ali, Ali Alqahtani, Mark W. Jones, and Xianghua Xie. “Clustering and Classification for Time Series Data in Visual Analytics: A Survey”. In: *IEEE Access* 7 (2019), pp. 181314–181338. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2958551](https://doi.org/10.1109/ACCESS.2019.2958551).
- [224] Jurgen Bernard, David Sessler, Jorn Kohlhammer, and Roy A. Ruddle. “Using Dashboard Networks to Visualize Multiple Patient Histories: A Design Study on Post-Operative Prostate Cancer”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.3 (2019), pp. 1615–1628. ISSN: 19410506. DOI: [10.1109/TVCG.2018.2803829](https://doi.org/10.1109/TVCG.2018.2803829).
- [225] Michele Bernardini, Luca Romeo, Paolo Misericordia, and Emanuele Frontoni. “Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine”. In: *IEEE Journal of Biomedical and Health Informatics* (2019), pp. 1–1. ISSN: 2168-2194. DOI: [10.1109/jbhi.2019.2899218](https://doi.org/10.1109/jbhi.2019.2899218).



- [226] Dennis Dingen, Marcel Van't Veer, Patrick Houthuizen, Eveline H.J. Mestrom, Erik H.H.M. Korsten, Arthur R.A. Bouwman, and Jarke Van Wijk. "Regression-Explorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 246–255. DOI: [10.1109/TVCG.2018.2865043](https://doi.org/10.1109/TVCG.2018.2865043).
- [227] Cong Feng, Minglun Gong, and Oliver Deussen. "Treemapping via Balanced Partitioning". In: *Proc. Computational Visual Media*. 2019.
- [228] Benjamin S. Glicksberg, Boris Oskotsky, Phyllis M. Thangaraj, Nicholas Giangreco, Marcus A. Badgeley, Kipp W. Johnson, Debajyoti Datta, Vivek A. Rudrapatna, Nadav Rappoport, Mark M. Shervev, Riccardo Miotto, Theodore C. Goldstein, Eugenia Rutenberg, Remi Frazier, Nelson Lee, Sharat Israni, Rick Larsen, Bethany Percha, Li Li, Joel T. Dudley, Nicholas P. Tatonetti, and Atul J. Butte. "PatientExploreR: An Extensible Application for Dynamic Visualization of Patient Clinical History from Electronic Health Records in the OMOP Common Data Model". In: *Bioinformatics (Oxford, England)* 35.21 (Nov. 2019). Ed. by Jonathan Wren, pp. 4515–4518. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btz409](https://doi.org/10.1093/bioinformatics/btz409).
- [229] Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, and Anastasia Bezirianou. "Comparing Similarity Perception in Time Series Visualizations". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (Jan. 2019), pp. 523–533. ISSN: 1941-0506. DOI: [10.1109/TVCG.2018.2865077](https://doi.org/10.1109/TVCG.2018.2865077).
- [230] Marc N. Gourevitch, Jessica K. Athens, Shoshanna E. Levine, Neil Kleiman, and Lorna E. Thorpe. "City-Level Measures of Health, Health Determinants, and Equity to Foster Population Health Improvement: The City Health Dashboard". In: *American Journal of Public Health* 109.4 (Apr. 2019), pp. 585–592. ISSN: 0090-0036. DOI: [10.2105/AJPH.2018.304903](https://doi.org/10.2105/AJPH.2018.304903).
- [231] Shunan Guo, Zhuochen Jin, David Gotz, Fan Du, Hongyuan Zha, and Nan Cao. "Visual Progression Analysis of Event Sequence Data". In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (Jan. 2019), pp. 417–426. ISSN: 1077-2626. DOI: [10.1109/TVCG.2018.2864885](https://doi.org/10.1109/TVCG.2018.2864885).
- [232] The Allen Institute for Artificial Intelligence. *Semantic Scholar - An Academic Search Engine for Scientific Articles*. <https://www.semanticscholar.org/>. 2019.
- [233] Astrik Jeitler, Alpin Türkoglu, Denis Makarov, Timo Jockers, Juri Buchmüller, Udo Schlegel, and Daniel A. Keim. "RescueMark: Visual Analytics of Social Media Data for Guiding Emergency Response in Disaster Situations: Award for Skillful Integration of Language Model". In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. Oct. 2019, pp. 120–121. DOI: [10.1109/VAST47406.2019.8986898](https://doi.org/10.1109/VAST47406.2019.8986898).
- [234] Kaveh Kamgar, Shaghayegh Gharghabi, and Eamonn Keogh. "Matrix Profile XV: Exploiting Time Series Consensus Motifs to Find Structure in Time Series Sets". In: *2019 IEEE International Conference on Data Mining (ICDM)*. Beijing, China: IEEE, Nov. 2019, pp. 1156–1161. ISBN: 978-1-7281-4604-1. DOI: [10.1109/ICDM.2019.00140](https://doi.org/10.1109/ICDM.2019.00140). (Visited on 05/12/2023).
- [235] Ellen Kim, Samuel M. Rubinstein, Kevin T. Nead, Andrzej P. Wojcieszynski, Peter E. Gabriel, and Jeremy L. Warner. "The Evolving Use of Electronic Health Records (EHR) for Research". In: *Seminars in Radiation Oncology* 29.4 (Oct. 2019), pp. 354–361. ISSN: 10534296. DOI: [10.1016/j.semradonc.2019.05.010](https://doi.org/10.1016/j.semradonc.2019.05.010).

- [236] Theresa A. Koleck, Caitlin Dreisbach, Philip E. Bourne, and Suzanne Bakken. “Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review”. In: *Journal of the American Medical Informatics Association* 26.4 (2019), pp. 364–379. ISSN: 1527974X. DOI: [10.1093/jamia/ocy173](https://doi.org/10.1093/jamia/ocy173).
- [237] Xiangjie Kong, Menglin Li, Gaoxing Zhao, Huijie Zhang, and Feng Xia. “COOC: Visual Exploration of Co-Occurrence Mobility Patterns in Urban Scenarios”. In: *IEEE Transactions on Computational Social Systems* 6.3 (June 2019), pp. 403–413. ISSN: 2329-924X, 2373-7476. DOI: [10.1109/TCSS.2018.2883582](https://doi.org/10.1109/TCSS.2018.2883582). (Visited on 10/27/2023).
- [238] Bum Chul Kwon, Min Je Choi, Joanne Taery Kim, Edward Choi, Young Bin Kim, Soonwook Kwon, Jimeng Sun, and Jaegul Choo. “RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records”. In: *IEEE Transactions on Visualization and Computer Graphics* (2019). ISSN: 19410506. DOI: [10.1109/TVCG.2018.2865027](https://doi.org/10.1109/TVCG.2018.2865027).
- [239] Sean Law. “STUMPY: A Powerful and Scalable Python Library for Time Series Data Mining”. In: *Journal of Open Source Software* 4.39 (July 2019), p. 1504. ISSN: 2475-9066. DOI: [10.21105/joss.01504](https://doi.org/10.21105/joss.01504). (Visited on 05/12/2023).
- [240] Jie Li, Siming Chen, Kang Zhang, Gennady Andrienko, and Natalia Andrienko. “COPE: Interactive Exploration of Co-Occurrence Patterns in Spatial Time Series”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.8 (Aug. 2019), pp. 2554–2567. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2018.2851227](https://doi.org/10.1109/TVCG.2018.2851227). (Visited on 10/26/2023).
- [241] Liam McNabb and Robert S. Laramee. “Multivariate Maps—A Glyph-Placement Algorithm to Support Multivariate Geospatial Visualization”. In: *Information* 10.10 (Sept. 2019), p. 302. ISSN: 2078-2489. DOI: [10.3390/info10100302](https://doi.org/10.3390/info10100302).
- [242] Huyen N. Nguyen and Tommy Dang. “EQSA: Earthquake Situational Analytics from Social Media”. In: *2019 IEEE Conference on Visual Analytics Science and Technology (VAST)*. Oct. 2019, pp. 142–143. DOI: [10.1109/VAST47406.2019.8986947](https://doi.org/10.1109/VAST47406.2019.8986947).
- [243] Florian Pappenberger, Hannah L. Cloke, and Calum A. Baugh. “Cartograms for Use in Forecasting Weather-Driven Natural Hazards”. In: *The Cartographic Journal* 56.2 (Apr. 2019), pp. 134–145. ISSN: 0008-7041, 1743-2774. DOI: [10.1080/00087041.2018.1534358](https://doi.org/10.1080/00087041.2018.1534358). (Visited on 03/03/2022).
- [244] Duccio Rocchini, Matteo Marcantonio, George Arhonditsis, Alessandro Lo Casciato, Heidi C. Hauffe, and Kate S. He. “Cartogramming Uncertainty in Species Distribution Models: A Bayesian Approach”. In: *Ecological Complexity* 38 (Apr. 2019), pp. 146–155. ISSN: 1476945X. DOI: [10.1016/j.ecocom.2019.04.002](https://doi.org/10.1016/j.ecocom.2019.04.002). (Visited on 03/03/2022).
- [245] James M. Salter, Daniel B. Williamson, John Scinocca, and Viatcheslav Kharin. “Uncertainty Quantification for Computer Models With Spatial Output Using Calibration-Optimal Bases”. In: *Journal of the American Statistical Association* 114.528 (Oct. 2019), pp. 1800–1814. ISSN: 0162-1459, 1537-274X. DOI: [10.1080/01621459.2018.1514306](https://doi.org/10.1080/01621459.2018.1514306). (Visited on 07/14/2024).
- [246] Harvard Medical School. *N2C2: National NLP Clinical Challenges*. <https://n2c2.dbmi.hms.harvard.edu/>. 2019.

- [247] Ben Shneiderman and Catherine Plaisant. “Interactive Visual Event Analytics: Opportunities and Challenges”. In: *Computer* 52.1 (Jan. 2019), pp. 27–35. ISSN: 0018-9162. DOI: [10.1109/MC.2018.2890217](https://doi.org/10.1109/MC.2018.2890217).
- [248] Nicole Sultanum, Devin Singh, Michael Brudno, and Fanny Chevalier. “Docurate: A Curation-Based Approach for Clinical Text Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 142–151. ISSN: 19410506. DOI: [10.1109/TVCG.2018.2864905](https://doi.org/10.1109/TVCG.2018.2864905).
- [249] Health Data Research UK. *Health Data Research Innovation Gateway*. <https://healthdatagateway.org/>. 2019.
- [250] Xudong Zhang, Jiehao Xiao, and Feng Gu. “Applying Support Vector Machine to Electronic Health Records for Cancer Classification”. In: 2019 Spring Simulation Conference, SpringSim 2019. IEEE, Apr. 2019, pp. 1–9. ISBN: 978-1-5108-8388-8. DOI: [10.23919/SpringSim.2019.8732906](https://doi.org/10.23919/SpringSim.2019.8732906).
- [251] Yixuan Zhang, Kartik Chanana, and Cody Dunne. “IDMVis: Temporal Event Sequence Visualization for Type 1 Diabetes Treatment Decision Support”. In: *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), pp. 512–522. ISSN: 19410506. DOI: [10.1109/TVCG.2018.2865076](https://doi.org/10.1109/TVCG.2018.2865076).
- [252] Fati Chen, Laurent Piccinini, Pascal Poncelet, and Arnaud Sallaberry. “Node Overlap Removal Algorithms: An Extended Comparative Study”. In: *Journal of Graph Algorithms and Applications* 24.4 (2020), pp. 683–706. ISSN: 1526-1719. DOI: [10.7155/jgaa.00532](https://doi.org/10.7155/jgaa.00532). (Visited on 06/17/2021).
- [253] Jian Chen, Meng Ling, Rui Li, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Torsten Moller, Robert Laramee Robert, Han-Wei Shen, Katherine Wünsche, and Qiru Wang. *IEEE Vis Figures and Tables Image Dataset*. July 2020. DOI: [10.21227/4HY6-VH52](https://doi.org/10.21227/4HY6-VH52). (Visited on 08/29/2022).
- [254] M. Chen, A. Abdul-Rahman, D. Archambault, J. Dykes, A. Slingsby, P. D. Ritsos, T. Torsney-Weir, C. Turkay, B. Bach, A. Brett, H. Fang, R. Jianu, S. Khan, R. S. Laramee, P. H. Nguyen, R. Reeve, J. C. Roberts, F. Vidal, Q. Wang, J. Wood, and K. Xu. “RAMPVIS: Towards a New Methodology for Developing Visualisation Capabilities for Large-scale Emergency Responses”. In: (Dec. 2020). DOI: [10.48550/ARXIV.2012.04757](https://doi.org/10.48550/ARXIV.2012.04757). (Visited on 08/29/2022).
- [255] Matteo Chinazzi, Jessica T. Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kungpeng Mu, Luca Rossi, Kaiyuan Sun, Cécile Viboud, Xinyue Xiong, Hongjie Yu, M. Elizabeth Halloran, Ira M. Longini, and Alessandro Vespignani. “The Effect of Travel Restrictions on the Spread of the 2019 Novel Coronavirus (COVID-19) Outbreak”. In: *Science* 368.6489 (Apr. 2020), pp. 395–400. DOI: [10.1126/science.aba9757](https://doi.org/10.1126/science.aba9757). (Visited on 06/01/2023).
- [256] The Scottish COVID-19 Response Consortium. *The Scottish COVID-19 Response Consortium (SCRC)*. <https://www.gla.ac.uk/research/az/scrc>. 2020.
- [257] Ensheng Dong, Hongru Du, and Lauren Gardner. “An Interactive Web-Based Dashboard to Track COVID-19 in Real Time”. In: *The Lancet Infectious Diseases* (Feb. 2020). ISSN: 14733099. DOI: [10.1016/S1473-3099\(20\)30120-1](https://doi.org/10.1016/S1473-3099(20)30120-1).
- [258] Mai Elshehaly, Rebecca Randell, Matthew Brehmer, Lynn McVey, Natasha Alvarado, Chris P. Gale, and Roy A. Ruddell. “QualDash: Adaptable Generation of Visualisation Dashboards for Healthcare Quality Improvement”. In: *IEEE Transactions on Visualization and Computer Graphics* (2020), pp. 1–1. ISSN: 1077-2626. DOI: [10.1109/TVCG.2020.3030424](https://doi.org/10.1109/TVCG.2020.3030424).



- [259] Peichao Gao, Hong Zhang, Zhiwei Wu, and Jicheng Wang. “Visualising the Expansion and Spread of Coronavirus Disease 2019 by Cartograms”. In: *Environment and Planning A: Economy and Space* 52.4 (June 2020), pp. 698–701. ISSN: 0308-518X, 1472-3409. DOI: [10.1177/0308518X20910162](https://doi.org/10.1177/0308518X20910162). (Visited on 12/08/2021).
- [260] Alberto Godio, Francesca Pace, and Andrea Vergnano. “SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence”. In: *International Journal of Environmental Research and Public Health* 17.10 (Jan. 2020), p. 3535. ISSN: 1660-4601. DOI: [10.3390/ijerph17103535](https://doi.org/10.3390/ijerph17103535). (Visited on 06/01/2023).
- [261] David Gotz, Jonathan Zhang, Wenyuan Wang, Joshua Shrestha, and David Borland. “Visual Analysis of High-Dimensional Event Sequence Data via Dynamic Hierarchical Aggregation”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.1 (Jan. 2020), pp. 440–450. ISSN: 1941-0506. DOI: [10.1109/TVCG.2019.2934661](https://doi.org/10.1109/TVCG.2019.2934661).
- [262] Shaobo He, Yuexi Peng, and Kehui Sun. “SEIR Modeling of the COVID-19 and Its Dynamics”. In: *Nonlinear Dynamics* 101.3 (Aug. 2020), pp. 1667–1680. ISSN: 0924-090X, 1573-269X. DOI: [10.1007/s11071-020-05743-y](https://doi.org/10.1007/s11071-020-05743-y). (Visited on 06/01/2023).
- [263] The Allen Institute for Artificial Intelligence. *COVID-19 Open Research Dataset Challenge (CORD-19)* — Kaggle. <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/kernels>. 2020.
- [264] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. “CarePre: An Intelligent Clinical Decision Assistance System”. In: *ACM Transactions on Computing for Healthcare* 1.1 (Mar. 2020), pp. 1–20. ISSN: 2691-1957. DOI: [10.1145/3344258](https://doi.org/10.1145/3344258).
- [265] Bum Chul Kwon, Vibha Anand, Kristen A. Severson, Soumya Ghosh, Zhaonan Sun, Brigitte I. Frohnert, Markus Lundgren, and Kenney Ng. “DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.26.c (2020), pp. 1–1. ISSN: 1077-2626. DOI: [10.1109/TVCG.2020.2985689](https://doi.org/10.1109/TVCG.2020.2985689).
- [266] Allison McCartney, Brittany Harris, Mira Rojanasakul, Julian Burgess, Paul Murray, Alyssa Vann, Demetrios Pogkas, Brad Benhamou, Aaron Kessler, and Alex Tribou. “2020 Presidential Election Results: Live Updates”. In: *Bloomberg* (2020). <https://www.bloomberg.com/graphics/2020-us-election-results>. (Visited on 01/16/2022).
- [267] Netlify. *Coronavirus/COVID-19 Support - Netlify*. <https://www.netlify.com/blog/2020/03/22/coronavirus/covid-19-support/>. Mar. 2020. (Visited on 02/24/2024).
- [268] Sabrina Nusrat, Jawaherul Alam, and Stephen Kobourov. “Recognition and Recall of Geographic Data In Cartograms”. In: *Proceedings of the International Conference on Advanced Visual Interfaces*. Salerno Italy: ACM, Sept. 2020, pp. 1–9. ISBN: 978-1-4503-7535-1. DOI: [10.1145/3399715.3399873](https://doi.org/10.1145/3399715.3399873). (Visited on 12/08/2021).
- [269] Seula Park, Gunhak Lee, and Jung Ok Kim. “Flood Evacuation Mapping Using a Time–Distance Cartogram”. In: *ISPRS International Journal of Geo-Information* 9.4 (Mar. 2020), p. 207. ISSN: 2220-9964. DOI: [10.3390/ijgi9040207](https://doi.org/10.3390/ijgi9040207). (Visited on 03/03/2022).

- [270] Bernhard Preim and Kai Lawonn. “A Survey of Visual Analytics for Public Health”. In: *Computer Graphics Forum* 39.1 (Feb. 2020), pp. 543–580. ISSN: 0167-7055. DOI: [10.1111/cgf.13891](https://doi.org/10.1111/cgf.13891).
- [271] *Rapid Assistance in Modelling the Pandemic: RAMP — Royal Society*. <https://royalsociety.org/news-resources/projects/ramp/>. 2020. (Visited on 05/13/2023).
- [272] Dylan Rees, Qiru Wang, and Robert S. Laramee. “The Industry Engagement Ladder”. In: *Journal of Industry-University Collaboration* 2.3 (Aug. 2020), pp. 125–139. ISSN: 2631-357X. DOI: [10.1108/JIUC-02-2020-0001](https://doi.org/10.1108/JIUC-02-2020-0001). (Visited on 06/26/2021).
- [273] Willy Scheibel, Daniel Limberger, and Jürgen Döllner. “Survey of Treemap Layout Algorithms”. In: *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*. Eindhoven Netherlands: ACM, Dec. 2020, pp. 1–9. ISBN: 978-1-4503-8750-7. DOI: [10.1145/3430036.3430041](https://doi.org/10.1145/3430036.3430041). (Visited on 07/10/2022).
- [274] Observational Health Data Sciences and Informatics. *The Book of OHDSI*. 1st ed. <http://book.ohdsi.org>. 2020. ISBN: 978-1-0888-5519-5.
- [275] Scottish COVID-19 Response Consortium. *Covid19\_EERAModel*. [https://github.com/ScottishCovidResponse/Covid19\\_EERAModel](https://github.com/ScottishCovidResponse/Covid19_EERAModel). Sept. 2020. (Visited on 05/25/2023).
- [276] The Learning Network. “What’s Going On in This Graph? — 2020 Presidential Election Maps”. In: *The New York Times* (Nov. 2020). <https://www.nytimes.com/2020/11/19/learning/whats-going-on-in-this-graph-2020-presidential-election-maps.html>. ISSN: 0362-4331. (Visited on 01/16/2022).
- [277] Johns Hopkins University. *Coronavirus COVID-19 (2019-nCoV)*. <https://www.arcgis.com/apps/opsdashboard/index.html>. 2020.
- [278] Johns Hopkins University. *Novel Coronavirus (COVID-19) Cases, Provided by JHU CSSE*. <https://github.com/CSSEGISandData/COVID-19>. 2020.
- [279] *University of Glasgow - The Scottish COVID-19 Response Consortium*. <https://www.gla.ac.uk/research/az/scrc/>. 2020. (Visited on 05/13/2023).
- [280] Alfredo Vellido. “The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care”. In: *Neural Computing and Applications* 32.24 (Dec. 2020), pp. 18069–18083. ISSN: 0941-0643, 1433-3058. DOI: [10.1007/s00521-019-04051-w](https://doi.org/10.1007/s00521-019-04051-w). (Visited on 10/06/2023).
- [281] *Visualization and Visual Analytics in Support of Rapid Assistance in Modelling the Pandemic (RAMP)*. <https://sites.google.com/view/rampvis>. 2020. (Visited on 05/13/2023).
- [282] Sebastian Vollmer, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine S L McAllister, Puja Myles, David Grainger, Mark Birse, Richard Branson, Karel G M Moons, Gary S Collins, John P A Ioannidis, Chris Holmes, and Harry Hemingway. “Machine Learning and Artificial Intelligence Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics, and Effectiveness”. In: *BMJ* (Mar. 2020), p. l6927. ISSN: 1756-1833. DOI: [10.1136/bmj.l6927](https://doi.org/10.1136/bmj.l6927). (Visited on 10/06/2023).
- [283] Yasmin Abdalla, Harriet Auty, Lisa Boden, Alys Brett, Min Chen, Ruth Dundas, Louise Matthews, Iain McKendrick, Dominic Mellor, and Richard Reeve. *Scottish COVID-19 Response Consortium Stakeholder Report*. Tech. rep. 2021.

- [284] Iuliia Alieva. “How American Media Framed 2016 Presidential Election Using Data Visualization: The Case Study of the New York Times and the Washington Post”. In: *Journalism Practice* (May 2021), pp. 1–27. ISSN: 1751-2786, 1751-2794. DOI: [10.1080/17512786.2021.1930573](https://doi.org/10.1080/17512786.2021.1930573). (Visited on 03/03/2022).
- [285] David Alvarez Castro and Alistair Ford. “3D Agent-Based Model of Pedestrian Movements for Simulating COVID-19 Transmission in University Students”. In: *ISPRS International Journal of Geo-Information* 10.8 (Aug. 2021), p. 509. ISSN: 2220-9964. DOI: [10.3390/ijgi10080509](https://doi.org/10.3390/ijgi10080509). (Visited on 06/01/2023).
- [286] Dario Antweiler, David Sessler, Sebastian Ginzler, and Jörn Kohlhammer. “Towards the Detection and Visual Analysis of COVID-19 Infection Clusters”. In: *EuroVis Workshop on Visual Analytics (EuroVA)* (2021), 5 pages. DOI: [10.2312/EUROVA.20211097](https://doi.org/10.2312/EUROVA.20211097). (Visited on 11/24/2023).
- [287] Aldo Arranz-López, Julio A Soria-Lara, and Amor Ariza-Álvarez. “An End-User Evaluation to Analyze the Effectiveness of Cartograms for Mapping Relative Non-Motorized Accessibility”. In: *Environment and Planning B: Urban Analytics and City Science* 48.9 (Nov. 2021), pp. 2880–2897. ISSN: 2399-8083, 2399-8091. DOI: [10.1177/2399808321991541](https://doi.org/10.1177/2399808321991541). (Visited on 03/03/2022).
- [288] T. Baumgartl, M. Petzold, M. Wunderlich, M. Hohn, D. Archambault, M. Lieser, A. Dalpke, S. Scheithauer, M. Marschollek, V. M. Eichel, N. T. Mutters, Highmed Consortium, and T. Von Landesberger. “In Search of Patient Zero: Visual Analytics of Pathogen Transmission Pathways in Hospitals”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (Feb. 2021), pp. 711–721. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2020.3030437](https://doi.org/10.1109/TVCG.2020.3030437). (Visited on 11/24/2023).
- [289] Jian Chen, Meng Ling, Rui Li, Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Torsten Moller, Robert S. Laramee, Han-Wei Shen, Katharina Wunsche, and Qiru Wang. “VIS30K: A Collection of Figures and Tables From IEEE Visualization Conference Publications”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.9 (Sept. 2021), pp. 3826–3833. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2021.3054916](https://doi.org/10.1109/TVCG.2021.3054916). (Visited on 09/30/2021).
- [290] Sara Di Bartolomeo, Yixuan Zhang, Fangfang Sheng, and Cody Dunne. “Sequence Braiding: Visual Overviews of Temporal Event Sequences and Attributes”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (Feb. 2021), pp. 1353–1363. ISSN: 1941-0506. DOI: [10.1109/TVCG.2020.3030442](https://doi.org/10.1109/TVCG.2020.3030442).
- [291] Engineering & Physical Sciences Research Council. *RAMP VIS: Making Visual Analytics an Integral Part of the Technological Infrastructure for Combating COVID-19*. <https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V054236/1>. Feb. 2021. (Visited on 06/09/2023).
- [292] Zhuochen Jin, Shunan Guo, Nan Chen, Daniel Weiskopf, David Gotz, and Nan Cao. “Visual Causality Analysis of Event Sequence Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.2 (Feb. 2021), pp. 1343–1352. ISSN: 1941-0506. DOI: [10.1109/TVCG.2020.3030465](https://doi.org/10.1109/TVCG.2020.3030465).
- [293] Xiaoxiao Liu, Mohammad Alharbi, Joe Best, Jian Chen, Alexandra Diehl, Elif Firat, Dylan Rees, Qiru Wang, and Robert S Laramee. “Visualization Resources: A Starting Point”. In: *The 25th International Conference on Information Visualization*. Sydney, Australia, 2021, pp. 160–169. ISBN: 978-1-6654-3827-8. DOI: [10.1109/IV53921.2021.00034](https://doi.org/10.1109/IV53921.2021.00034).

- [294] Daniel E. Sack, Stephen J. Gange, Keri N. Althoff, April C. Pettit, Asghar N. Kheshti, Imani S. Ransby, Jeff J. Nelson, Megan M. Turner, Timothy R. Sterling, and Peter F. Rebeiro. “Visualizing the Geography of HIV Observational Cohorts with Density-Adjusted Cartograms”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* Publish Ahead of Print (Dec. 2021). ISSN: 1525-4135. DOI: [10.1097/QAI.0000000000002903](https://doi.org/10.1097/QAI.0000000000002903). (Visited on 03/03/2022).
- [295] Sarah Schöttler, Yalong Yang, Hanspeter Pfister, and Benjamin Bach. “Visualizing and Interacting with Geospatial Networks: A Survey and Design Space”. In: *Computer Graphics Forum* 40.6 (Sept. 2021), pp. 5–33. ISSN: 0167-7055, 1467-8659. DOI: [10.1111/cgf.14198](https://doi.org/10.1111/cgf.14198). (Visited on 07/10/2022).
- [296] Julien Siebert, Janek Groß, and Christof Schroth. “A Systematic Review of Packages for Time Series Analysis”. In: *The 7th International Conference on Time Series and Forecasting*. MDPI, June 2021, p. 22. DOI: [10.3390/engproc2021005022](https://doi.org/10.3390/engproc2021005022). (Visited on 05/12/2023).
- [297] Qiru Wang. *Thevisgroup/EnsembleVis*. <https://github.com/thevisgroup/EnsembleVis>. Aug. 2021. (Visited on 03/15/2024).
- [298] Qiru Wang, Robert S. Laramée, Arron Lacey, and William Owen Pickrell. “LetterVis: A Letter-Space View of Clinic Letters”. In: *The Visual Computer* 37.9-11 (Sept. 2021), pp. 2643–2656. ISSN: 0178-2789, 1432-2315. DOI: [10.1007/s00371-021-02171-w](https://doi.org/10.1007/s00371-021-02171-w). (Visited on 09/30/2021).
- [299] G. J. Ackland, J. Panovska-Griffiths, W. Waites, and M. E. Cates. “The Royal Society RAMP Modelling Initiative”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2233 (Aug. 2022), p. 20210316. DOI: [10.1098/rsta.2021.0316](https://doi.org/10.1098/rsta.2021.0316). (Visited on 05/30/2023).
- [300] Mohammad Alharbi, Robert S Laramée, and Tom Cheesman. “TransVis: Integrated Distant and Close Reading of Othello Translations”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.2 (Feb. 2022), pp. 1397–1414. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2020.3012778](https://doi.org/10.1109/TVCG.2020.3012778). (Visited on 07/10/2022).
- [301] M. Chen, A. Abdul-Rahman, D. Archambault, J. Dykes, P.D. Ritsos, A. Slingsby, T. Torsney-Weir, C. Turkay, B. Bach, R. Borgo, A. Brett, H. Fang, R. Jianu, S. Khan, R.S. Laramée, L. Matthews, P.H. Nguyen, R. Reeve, J.C. Roberts, F.P. Vidal, Q. Wang, J. Wood, and K. Xu. “RAMPVIS: Answering the Challenges of Building Visualisation Capabilities for Large-Scale Emergency Responses”. In: *Epidemics* 39 (June 2022), p. 100569. ISSN: 17554365. DOI: [10.1016/j.epidem.2022.100569](https://doi.org/10.1016/j.epidem.2022.100569). (Visited on 06/02/2022).
- [302] Jason Dykes, Alfie Abdul-Rahman, Daniel Archambault, Benjamin Bach, Rita Borgo, Min Chen, Jessica Enright, Hui Fang, Elif E. Firat, Euan Freeman, Tuna Gönen, Claire Harris, Radu Jianu, Nigel W. John, Saiful Khan, Andrew Lahiff, Robert S. Laramée, Louise Matthews, Sibylle Mohr, Phong H. Nguyen, Alma A. M. Rahat, Richard Reeve, Panagiotis D. Ritsos, Jonathan C. Roberts, Aidan Slingsby, Ben Swallow, Thomas Torsney-Weir, Cagatay Turkay, Robert Turner, Franck P. Vidal, Qiru Wang, Jo Wood, and Kai Xu. “Visualization for Epidemiological Modelling: Challenges, Solutions, Reflections and Recommendations”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380.2233 (Oct. 2022), p. 20210299. ISSN: 1364-503X, 1471-2962. DOI: [10.1098/rsta.2021.0299](https://doi.org/10.1098/rsta.2021.0299). (Visited on 08/29/2022).

- [303] Shunan Guo, Zhuochen Jin, Qing Chen, David Gotz, Hongyuan Zha, and Nan Cao. “Interpretable Anomaly Detection in Event Sequences via Sequence Matching and Visual Comparison”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.12 (Dec. 2022), pp. 4531–4545. ISSN: 1941-0506. DOI: [10.1109/TVCG.2021.3093585](https://doi.org/10.1109/TVCG.2021.3093585).
- [304] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. “Survey on Visual Analysis of Event Sequence Data”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.12 (Dec. 2022), pp. 5091–5112. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2021.3100413](https://doi.org/10.1109/TVCG.2021.3100413). (Visited on 05/31/2023).
- [305] Saiful Khan, Phong H. Nguyen, Alfie Abdul-Rahman, Benjamin Bach, Min Chen, Euan Freeman, and Cagatay Turkay. “Propagating Visual Designs to Numerous Plots and Dashboards”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (Jan. 2022), pp. 86–95. ISSN: 1941-0506. DOI: [10.1109/TVCG.2021.3114828](https://doi.org/10.1109/TVCG.2021.3114828).
- [306] Saiful Khan, Phong Hai Nguyen, Alfie Abdul-Rahman, Euan Freeman, Cagatay Turkay, and Min Chen. “Rapid Development of a Data Visualization Service in an Emergency Response”. In: *IEEE Transactions on Services Computing* 15.3 (May 2022), pp. 1251–1264. ISSN: 1939-1374, 2372-0204. DOI: [10.1109/TSC.2022.3164146](https://doi.org/10.1109/TSC.2022.3164146). (Visited on 06/01/2023).
- [307] Yoshiki Kusunoki, Kosuke Konishi, Taku Tsunoda, and Hidenori Koyama. “Significance of Glycemic Variability in Diabetes Mellitus”. In: *Internal Medicine* 61.3 (Feb. 2022), pp. 281–290. ISSN: 0918-2918, 1349-7235. DOI: [10.2169/internalmmedicine.8424-21](https://doi.org/10.2169/internalmmedicine.8424-21). (Visited on 10/06/2023).
- [308] Jessica Magallanes, Tony Stone, Paul D Morris, Suzanne Mason, Steven Wood, and Maria-Cruz Villa-Uriol. “Sequen-C: A Multilevel Overview of Temporal Event Sequences”. In: *IEEE Transactions on Visualization and Computer Graphics* 28.1 (Jan. 2022), pp. 901–911. ISSN: 1941-0506. DOI: [10.1109/TVCG.2021.3114868](https://doi.org/10.1109/TVCG.2021.3114868).
- [309] Soeren Nickel, Max Sondag, Wouter Meulemans, Stephen G. Kobourov, Jaakko Peltonen, and Martin Nollenburg. “Multicriteria Optimization for Dynamic Demers Cartograms”. In: *IEEE Transactions on Visualization and Computer Graphics* (2022), pp. 1–1. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2022.3151227](https://doi.org/10.1109/TVCG.2022.3151227). (Visited on 03/03/2022).
- [310] Ben Swallow, Paul Birrell, Joshua Blake, Mark Burgman, Peter Challenor, Luc E. Coffeng, Philip Dawid, Daniela De Angelis, Michael Goldstein, Victoria Hemming, Glenn Marion, Trevelyan J. McKinley, Christopher E. Overton, Jasmina Panovska-Griffiths, Lorenzo Pellis, Will Probert, Katriona Shea, Daniel Villela, and Ian Vernon. “Challenges in Estimation, Uncertainty Quantification and Elicitation for Pandemic Modelling”. In: *Epidemics* 38 (Mar. 2022), p. 100547. ISSN: 17554365. DOI: [10.1016/j.epidem.2022.100547](https://doi.org/10.1016/j.epidem.2022.100547). (Visited on 05/30/2023).
- [311] Q. Wang and R.S. Laramée. “EHR STAR: The State-Of-the-Art in Interactive EHR Visualization”. In: *Computer Graphics Forum* 41.1 (Feb. 2022), pp. 69–105. ISSN: 0167-7055, 1467-8659. DOI: [10.1111/cgf.14424](https://doi.org/10.1111/cgf.14424). (Visited on 08/10/2024).



- [312] Zhong Wang, Yongguo Han, Guijuan Wang, Weixin Zhao, Jiansong Wang, and Yadong Wu. “MCC-Vis: Visual Analysis for City Regional Co-occurrence Pattern Based on Traffic Trajectory Data”. In: *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*. Chengdu, China: IEEE, Aug. 2022, pp. 1245–1254. ISBN: 978-1-6654-9916-3. DOI: [10.1109/PRAI55851.2022.9904131](https://doi.org/10.1109/PRAI55851.2022.9904131). (Visited on 10/27/2023).
- [313] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tomin-ski. *Visualization of Time-Oriented Data*. Human–Computer Interaction Series. London: Springer London, 2023. ISBN: 978-1-4471-7526-1 978-1-4471-7527-8. DOI: [10.1007/978-1-4471-7527-8](https://doi.org/10.1007/978-1-4471-7527-8). (Visited on 02/01/2024).
- [314] Zikun Deng, Shifu Chen, Tobias Schreck, Dazhen Deng, Tan Tang, Mingliang Xu, Di Weng, and Yingcai Wu. “Visualizing Large-Scale Spatial Time Series with GeoChron”. In: *IEEE Transactions on Visualization and Computer Graphics* (Oct. 2023). DOI: [10.1109/TVCG.2023.3327162](https://doi.org/10.1109/TVCG.2023.3327162).
- [315] Anna Scimone, Klaus Eckelt, Marc Streit, and Andreas Hinterreiter. “Marjorie: Visualizing Type 1 Diabetes Data to Support Pattern Exploration”. In: *IEEE Transactions on Visualization and Computer Graphics* (2023), pp. 1–11. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2023.3326936](https://doi.org/10.1109/TVCG.2023.3326936). (Visited on 11/20/2023).
- [316] Anton Yeshchenko and Jan Mendling. “A Survey of Approaches for Event Sequence Analysis and Visualization”. In: *Information Systems* (Sept. 2023), p. 102283. ISSN: 0306-4379. DOI: [10.1016/j.is.2023.102283](https://doi.org/10.1016/j.is.2023.102283). (Visited on 09/12/2023).
- [317] Md. Jamal Hossain, Md. Al-Mamun, and Md. Rabiul Islam. “Diabetes Mellitus, the Fastest Growing Global Public Health Concern: Early Detection Should Be Focused”. In: *Health Science Reports* 7.3 (Mar. 2024), e2004. ISSN: 2398-8835, 2398-8835. DOI: [10.1002/hsr2.2004](https://doi.org/10.1002/hsr2.2004). (Visited on 09/21/2024).
- [318] Qiru Wang, Rita Borgo, and Robert S Laramée. “EnsembleDashVis Views and Volunteers – A Retrospective and Early History”. In: *New Community Health Models*. Ed. by Marco Bassanello, Ruggero Geppini, Xin-Nong Li, and Amy Matecki. <https://doi.org/10.5772/intechopen.115029>. Rijeka: IntechOpen, Aug. 2024.
- [319] Qiru Wang, Kai Xu, and Robert S. Laramée. “Demers Cartogram with Rivers”. In: *Visual Informatics* (Sept. 2024), S2468502X24000445. ISSN: 2468502X. DOI: [10.1016/j.visinf.2024.09.003](https://doi.org/10.1016/j.visinf.2024.09.003). (Visited on 09/15/2024).
- [320] Susanne Zabel, Philipp Hennig, and Kay Nieselt. “VIPurPCA: Visualizing and Propagating Uncertainty in Principal Component Analysis”. In: *IEEE Transactions on Visualization and Computer Graphics* 30.4 (Apr. 2024), pp. 2011–2022. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2023.3345532](https://doi.org/10.1109/TVCG.2023.3345532). (Visited on 03/15/2024).
- [321] The U.S. General Services Administration. *Data.Gov*. <https://www.data.gov/>.
- [322] Lifelines Biobank. *Lifelines Biobank*. <https://www.lifelines.nl/>.
- [323] UK Biobank. *UK Biobank*. <https://www.ukbiobank.ac.uk/>.
- [324] Matthew Bloch. *Mapshaper*. <https://mapshaper.org/>. (Visited on 01/23/2022).
- [325] NHS Commissioning Board. *NHS England Data Catalogue*. <https://data.england.nhs.uk/>.
- [326] Department of Epidemiology Erasmus University Medical Center. *The Rotterdam Study*. <http://www.erasmus-epidemiology.nl/research/ergo.htm>.

- [327] Big Cities Health Coalition. *Data Platform — Big Cities Health Coalition*. <https://www.bigcitieshealth.org/city-data/>.
- [328] *D3 by Observable — The JavaScript Library for Bespoke Data Visualization*. <https://d3js.org/>. (Visited on 02/24/2024).
- [329] NHS Scotland Open Data. *Datasets - NHS Scotland Open Data*. <https://www.opendata.nhs.scot/dataset>.
- [330] NHS Digital. *Personal Health Records Definition*. <https://digital.nhs.uk/services/personal-health-records-adoption-service/personal-health-records-adoption-toolkit/initiating-a-personal-health-record/personal-health-records-definition>.
- [331] Elsevier. *Mendeley - Reference Management Software & Researcher Network*. <https://www.mendeley.com/>.
- [332] Public Health England. *PHE Data and Analysis Tools - GOV.UK*. <https://www.gov.uk/guidance/phe-data-and-analysis-tools>.
- [333] FAIRsharing. *FAIRsharing*. <https://fairsharing.org/>.
- [334] GIAN TT. *GIAN TT — Groningen Initiative to Analyse Type 2 Diabetes Treatment*. <https://www.giantt.nl/>.
- [335] Google. *Dataset Search*. <https://datasetsearch.research.google.com/>.
- [336] Google. *Google*. <https://www.google.com/>.
- [337] Google. *Google Scholar*. <https://scholar.google.com/>.
- [338] HealthData.gov. *HealthData.Gov*. <https://healthdata.gov/>.
- [339] *Home*. <https://www.snomed.org>. (Visited on 07/14/2024).
- [340] *ICD-10 Version:2019*. <https://icd.who.int/browse10/2019/en>. (Visited on 07/14/2024).
- [341] IEEE. *IEEE Xplore Digital Library*. <https://ieeexplore.ieee.org/Xplore/home.jsp>.
- [342] National Cancer Institute. *Definition of Electronic Health Record*. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/electronic-health-record>.
- [343] National Cancer Institute. *Definition of Personal Health Record*. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/personal-health-record>.
- [344] National Cancer Institute. *Surveillance, Epidemiology, and End Results Program*. <https://seer.cancer.gov/>.
- [345] *ISARIC 4C*. <https://isaric4c.net>. (Visited on 07/17/2022).
- [346] Johns Hopkins University. *COVID-19 Map*. <https://coronavirus.jhu.edu/map.html>. (Visited on 06/01/2023).
- [347] The Association for Computing Machinery. *ACM Digital Library*. <https://dl.acm.org/>.
- [348] Götzfried Antique Maps. *An Alphabetically, British Isles, England and Wales, 1679*. <https://www.vintage-maps.com/en/antique-maps/europe/british-isles/adams-british-isles-england-and-wales-1679-1699::12435>. (Visited on 07/03/2022).
- [349] The U.S. Centers for Medicare and Medicaid Services. *Electronic Health Records*. <https://www.cms.gov/Medicare/E-Health/EHealthRecords>.

- [350] NHS. *NHS Clinical Commissioning Groups (CCGs)*. <https://www.england.nhs.uk/commissioning/who-commissions-nhs-services/>. (Visited on 01/23/2022).
- [351] NHS Digital. *Clinical Commissioning Group Outcomes Indicator Set (CCG OIS)*. <https://digital.nhs.uk/data-and-information/publications/statistical/ccg-outcomes-indicator-set>. (Visited on 01/23/2022).
- [352] Open Geography portalx. *Open Geography Portalx*. <https://geoportal.statistics.gov.uk/datasets/d6acd30ad71f4e14b4de808e58d9bc4c>. (Visited on 01/23/2022).
- [353] OpenStreetMap. *Relation: Thames (2263653)*. <https://www.openstreetmap.org/relation/2263653>. (Visited on 01/23/2022).
- [354] Overpass Turbo. *Overpass Turbo*. <https://overpass-turbo.eu/>. (Visited on 01/23/2022).
- [355] QGIS. *Welcome to the QGIS Project!* <https://qgis.org/en/site/>. (Visited on 01/23/2022).
- [356] re3data.org. *Registry of Research Data Repositories*. DOI: 10.17616/R3D.
- [357] *Read Codes*. <https://digital.nhs.uk/services/terminology-and-classifications/read-codes>. (Visited on 07/14/2024).
- [358] Maelstrom Research. *Maelstrom Catalogue*. <https://www.maelstrom-research.org/maelstrom-catalogue>.
- [359] ResearchGate. *Search Publications — ResearchGate*. <https://www.researchgate.net/search/publications>.
- [360] The Government Digital Service. *Find Open Data - Data.Gov.Uk*. <https://data.gov.uk/>.
- [361] Vladimir Shkolnikov, Magali Barbieri, and John Wilmoth. *Human Mortality Database*. <https://www.mortality.org/>.
- [362] The Open Data Team. *Open Data NI*. <https://www.opendatani.gov.uk/>.
- [363] The Office of the National Coordinator for Health Information Technology. *Health IT Data*. <https://dashboard.healthit.gov/datadashboard/data.php>.
- [364] The Apache Software Foundation. *Apache cTAKES™ - Clinical Text Analysis Knowledge Extraction System*. <https://ctakes.apache.org/>. (Visited on 07/14/2024).
- [365] TopoJSON. *The TopoJSON Format Specification*. <https://github.com/topojson/topojson-specification>. (Visited on 01/23/2022).
- [366] Trails. *Tracking Adolescents' Individual Lives Survey*. <https://www.trails.nl/en>.
- [367] European Union. *European Data Portal*. <https://www.europeandataportal.eu/en>.
- [368] Public Health Wales. *Data - Public Health Wales*. <https://phw.nhs.wales/data/>.



# Appendices



# Appendix A

## Damon Berridge

This appendix is dedicated to the late Professor Damon Berridge.

As an expert in large healthcare data, Professor Berridge's work greatly inspired this thesis. He served as the Co-Investigator of the EP/S010238/1 EPSRC grant, which supported this Ph.D.



# Appendix B

## List of Domain Experts

Here we present a list of domain experts and collaborators with whom we have collaborated or consulted during the course of this thesis. The list is not exhaustive, but it provides a starting point for interested researchers to explore the field of EHR Vis and related disciplines.

### B.1 Alfie Abdul-Rahman

Dr Alfie Abdul-Rahman is a Senior Lecturer in Computer Science at King's College London. Her research interests include information visualisation, computer graphics, human-computer interaction, and digital humanities.

### B.2 Sara Di Bartolomeo

Dr Sara Di Bartolomeo is a postdoc researcher at Vienna University of Technology. Her research focuses on Graph Drawing - especially layered graphs, Generative Models and Virtual Reality.

### B.3 Rita Borgo

Dr Rita Borgo is the head of the Human Centred Computing Group at King's College London. Her research interests include Data Science, Augmented AI, Scientific and Information visualisation, Time Series analysis, Large Data sets and High Performance

Computing.

## **B.4 Peter Challenor**

Peter Challenor is a professor in the statistical sciences group of the Department of Mathematics of the University of Exeter. His research research is concerned with the development of methods to quantify the uncertainty for models of the environment, climate, engineering, and healthcare.

## **B.5 Min Chen**

Min Chen is a professor of scientific visualisation at Oxford University and a fellow of Pembroke College. His research interests include data visualisation, data science, computer graphics, computer vision, and human-computer interaction.

## **B.6 Alena Denisova**

Dr Alena Denisova is a senior lecturer at University of York. Her research interests include understanding and improving player experience of interactive media.

## **B.7 Cody Dunne**

Dr Cody Dunne is an associate professor at Northeastern University's Khoury College of Computer and Sciences. He works at the intersection of information visualisation, network science, human-computer interaction, and computer science. Dunne focuses on techniques for making data easier to analyse and share, as well as the application of visualisation techniques to real-world problems.

## **B.8 Arron Lacey**

Dr Arron Lacey is a lecturer in Health Data Science and Natural Language Processing at the Swansea University Medical School. Dr Lacey has been based in the SAIL Databank

at the Medical School since 2011 following a BSc in Physics, MScs in Computer Science and a Ph.D. in Healthcare Studies at Swansea University.

## **B.9 Ian Litchfield**

Ian Litchfield is a Research Fellow in the Institute of Applied Health Research, University of Birmingham, with a background in occupational medicine and interests in health service delivery. He uses qualitative and quantitative research methods in the evaluation of service delivery interventions in both primary and secondary care.

## **B.10 Owen Pickrell**

Dr Owen Pickrell is a consultant neurologist and honorary clinical associate professor at Swansea University Medical School. He practises clinically as a neurologist in Swansea Bay University Health board with a sub-speciality interest in epilepsy.

## **B.11 Panagiotis D. Ritsos**

Dr Panagiotis Ritsos is a Senior Lecturer (Associate Professor) in Visualisation, and the Director of Research at the School of Computer Science and Engineering, Bangor University. His research interests revolve around the domain of human-computer interaction (HCI) and include mixed/augmented and virtual reality (XR), information visualisation (InfoVis), visual analytics (VA) and wearable computing (WearComp).

## **B.12 Benjamin Swallow**

Dr Benjamin Swallow is a Lecturer in Statistics, School of Mathematics and Statistics, University of St Andrews. Ben's research interests lie largely in statistical inference in complex dynamic systems, often those changing in space and/or time.

## **B.13 Thomas Torsney-Weir**

Dr Thomas Torsney-Weir was a Lecturer in Computer Science at Swansea University when he provided his domain expertise to this Ph.D. He is a computer scientist with extensive experience building systems from the ground up as well as research in ML/AI with a focus on visual analysis in multidimensional continuous spaces and understanding complex models.

## **B.14 Cagatay Turkey**

Professor Cagatay Turkey from University of Warwick, his research falls under the broad area that can be referred to as Visual Data Science and focuses on designing visualisations, interactions and computational methods to enable an effective combination of human and machine capabilities to facilitate data-intensive problem solving.

## **B.15 Samantha Turner**

Samantha Turner is a Research Officer and Data Scientist specialising in the field of Injury Prevention. Samantha has worked on several data linkage and research projects with the aim to improve injury profiling, injury incidence estimates, the comparability of injury data across countries, and measurement of injury burden.

## **B.16 Ian Vernon**

Professor Ian Vernon is a statistician at Durham University. His research focuses on uncertainty quantification for galaxy formation, epidemiology, systems biology, geology, and nuclear physics.

## **B.17 Franck P. Vidal**

Dr Franck P. Vidal was a Lecturer in Computer Science at Bangor University when he provided his domain expertise to this Ph.D. His research area is mainly focusing on



computer graphics, visualisation and physically-based simulation for medical applications.

## **B.18 Phil Weber**

Dr Phil Weber is a lecturer in computer science and member of the Aston Centre for Artificial Intelligence Research and Applications (ACAIRA), and Aston Institute for Forensic Linguistics (AIFL), specifically the Forensic Data Science Laboratory (FDSL). His current research is in forensic voice comparison (which uses state of the art tools developed for automatic speaker recognition).

## **B.19 Kai Xu**

Dr Kai Xu is an Associate Professor in the School of Computer Science at the University of Nottingham. His main research interest is Data Science, particularly Data Visualisation, i.e., presenting data visually to facilitate pattern discovery using human cognition and domain knowledge. Kai is also the second supervisor of this Ph.D.



# Appendix C

## List of Data sets Used

Here we list a table of all data sets we have explored during this Ph.D.

Chapter	Description	Link
<a href="#">Chapter 3</a>	Epilepsy Clinic Letters	Supplied by Domain Experts
<a href="#">Chapter 4</a>	River Shapefiles	<a href="https://overpass-turbo.eu/">https://overpass-turbo.eu/</a>
<a href="#">Chapter 4</a>	NHS CCG Shapefiles	<a href="https://geoportal.statistics.gov.uk/maps/2f226df77c444d93aeebe5220cd50186">https://geoportal.statistics.gov.uk/maps/2f226df77c444d93aeebe5220cd50186</a>
<a href="#">Chapter 4</a>	NHS CCG Outcomes Indicator	<a href="https://digital.nhs.uk/data-and-information/publications/statistical/ccg-outcomes-indicator-set">https://digital.nhs.uk/data-and-information/publications/statistical/ccg-outcomes-indicator-set</a>
<a href="#">Chapter 6</a>	COVID-19 Outbreak Data in Scotland	<a href="https://github.com/ScottishCovidResponse">https://github.com/ScottishCovidResponse</a>
<a href="#">Chapter 6</a>	COVID-19 Scottish data modelling and simulations	<a href="https://github.com/thevisgroup/EnsembleVis">https://github.com/thevisgroup/EnsembleVis</a>
<a href="#">Chapter 5</a>	Glucose readings from continuous glucose monitors, food intake record, insulin dosage	<a href="https://github.com/VisDunneRight/IDMVis">https://github.com/VisDunneRight/IDMVis</a>
<a href="#">Chapter 5</a>	Glucose readings from continuous glucose monitors	<a href="https://github.com/jku-vds-lab/marjorie/">https://github.com/jku-vds-lab/marjorie/</a>

Table C.1: A list of data sets explored in this Ph.D.



## Appendix D

### List of Data Set Access Applications

Access to EHR data sets is a major challenge, as stated in [Section 2.6](#). Throughout this Ph.D., we applied for access to multiple EHR data sets, here we record the timeline of these applications. Due to the difficulty in obtaining access to these data sets, most of the work in this Ph.D. leveraged open access data sets, stated in [Section 2.6.5](#).

Name	Region	Description	Application Time	Application Outcome	Emails Exchanged
Epilepsy Clinic Letters	UK	The clinic letters are directly supplied by a group of researchers in Swansea University Medical School as part of a collaboration.	15 July 2020	200 letters granted in Oct 2022	> 30
ISARIC4C <a href="#">[345]</a>	UK	The ISARIC4C study has created an open-access integrated analysis platform for linked clinical data across the NHS for various studies.	17 Mar 2022	Under review	17
MIMIC-III <a href="#">[150]</a>	US	MIMIC-III (Medical Information Mart for Intensive Care) is a large, single-centre database comprising information relating to patients admitted to critical care units at a large tertiary care hospital.	14 May 2022	31 May 2022	4

Table D.1: A list of EHR data set access applications during this Ph.D. This table demonstrates the difficulty in accessing EHR data sets for research purposes.



# Appendix E

## List of Videos

During this Ph.D., multiple videos were created to illustrate the work conducted. The following links to these videos offer interested readers a clearer understanding of the research presented in this thesis.

Chapter	Description	Link
<a href="#">Chapter 2</a>	EHR STAR Presentation at the EuroVis 2022 Conference	<a href="https://youtu.be/8phSEunqdpw">https://youtu.be/8phSEunqdpw</a>
<a href="#">Chapter 3</a>	Demonstration for LetterVis	<a href="https://youtu.be/jSVzhCjLi_U">https://youtu.be/jSVzhCjLi_U</a>
<a href="#">Chapter 4</a>	Demonstration for Demers Cartogram with Rivers	<a href="https://youtu.be/DgCwCkyfGKk">https://youtu.be/DgCwCkyfGKk</a>
<a href="#">Chapter 4</a>	Demers Cartogram with Rivers User Study	<a href="https://www.youtube.com/playlist?list=PLL7sHvxLtD75fMtrUQrAdddj3wFkcWz">https://www.youtube.com/playlist?list=PLL7sHvxLtD75fMtrUQrAdddj3wFkcWz</a>
<a href="#">Chapter 5</a>	Demonstration and Case Studies for Time Series Map	<a href="https://youtu.be/TnlyZDQCpQE">https://youtu.be/TnlyZDQCpQE</a>

Table E.1: A collection of videos showcasing the research conducted during this Ph.D.