Commit2Data

Edited by

Boudewijn R. Haverkort
Aldert de Jongste
Pieter van Kuilenburg
Ruben D. Vromans

OASICS

Editors

Boudewijn R. Haverkort
Tilburg University, Tilburg, The Netherlands
B.R.H.M.Haverkort@tilburguniversity.edu

Aldert de Jongste
ECP, The Hague, The Netherlands
aldert.de.jongste@ecp.nl

Pieter van Kuilenburg
ECP, The Hague, The Netherlands
Pieter.vanKuilenburg@ecp.nl

Ruben D. Vromans
Tilburg University, Tilburg, The Netherlands
R.D.Vromans@tilburguniversity.edu

ACM Classification 2012
Information systems; Applied computing

ISBN 978-3-95977-351-5

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/978-3-95977-351-5.

Publication date
October, 2024

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://portal.dnb.de.

Digital Object Identifier: 10.4230/OASIcs.Commit2Data.2024.0

ISBN 978-3-95977-351-5 **ISSN 1868-8969** **https://www.dagstuhl.de/oasics**

OASIcs – OpenAccess Series in Informatics

OASIcs is a series of high-quality conference proceedings across all fields in informatics. OASIcs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

ISSN 1868-8969

https://www.dagstuhl.de/oasics

Contents

Regular Papers

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans
OASICS OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Preface

Commit2Data is the Dutch research and innovation program on the topic of big data analytics and applications; it was executed in the period 2016–2024. The program not only addresses use-inspired research on big data and applications, it also specifically addresses valorization and dissemination of results.

This book presents some highlights of the Commit2Data (C2D) program, with emphasis on its research projects. In the book a selection of eight research projects is presented, together with a description of the program as a whole, thus nicely giving an overview of the broadness of the program. We are grateful that this book is published by Dagstuhl Publishing, in its Open Access Series in Informatics (OASIcs).

The papers in this book were selected after an open call for contributions that was issued in the summer of 2022. We specifically called for project overview papers, that is, not papers that describe a single technical aspect, but papers that describe the whole set-up of a project, enhanced with some technical highlights. The call was open to projects outside of the Commit2Data program, however, it did not attract such contributions. In the cause of 2022, some projects within Commit2Data were already finished, others were still ongoing; it turned out difficult to attract contributions from projects that were already finished. After a review, selection and revision process that took place in 2023 and for which we used the EasyChair platform, we are now happy to present this selection of project descriptions.

The opening paper "The Public-Private Research and Innovation Program Commit2Data" by Haverkort *et al.* presents the overall set-up of the Commit2Data program, with ample attention for its public-private character and the involvement of many different application sectors. Specific attention is paid to the "C2D approach" to valorize research results, as well as to lessons learned from the program as a whole.

The subsequent three papers then address the health and vitality domain. In "Building a Digital Health Twin for Personalized Intervention: the EPI project", Alsayed Kassem *et al.* discuss their EPI (Enabling Personalized Interventions) project. The authors integrated research on data science, software-defined network infrastructure, and secure and trustworthy data sharing in the healthcare domain, using the digital twin paradigm. They also point to the challenges including strict data sharing policies, complex legal and privacy requirements, and infrastructure limitations, and discuss their solutions.

In the paper "Helping Cancer Patients to Choose the Best Treatment: Towards Automated Data-Driven and Personalized Information Presentation of Cancer Treatment Options", Krahmer *et al.* describe their interdisciplinary approach to automatically generate personalized treatment descriptions for cancer patients using data from the Netherlands Cancer Registry and the (so-called) PROFILES dataset. By presenting this information in verbal, numerical, and narrative formats, the project facilitates shared decision-making between doctors and patients about treatments. The paper discusses the strengths, limitations, and broader applicability of this approach.

In the paper "SeNiors empOWred via Big Data to Joint-Manage Their Medication-Related Risk of Falling in Primary Care: The SNOWDROP Project", Westerbeek *et al.* give an overview of the SNOWDROP project. This project developed and evaluated data-driven methods to predict individualized fall risks for elderly people. By creating validated prediction models from electronic health records, a clinical decision support system, and a patient portal, the project seeks to improve shared decision-making, medication management, and patient outcomes.

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans
OASICS OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

The paper "Real-time Data-Driven Maintenance Logistics: a Public-Private Collaboration" by Van Jaarsveld *et al.* addresses real-time data-driven maintenance logistics. The project aimed to innovate maintenance logistics through data-driven decision making by collaborating with three innovative companies and researchers from two leading knowledge institutions. The authors review innovations inspired by practice, their materialization, and their impact on practice.

Within the theme of sports, the paper "WheelPower: Wheelchair Sports and Data Science Push It to the Limit" by Vegter *et al.* provides a comprehensive overview of their project, which aims to improve the power output of athletes in sport-specific wheelchairs for better competition performance. The project brought together data scientists and all major Paralympic wheelchair sports, and integrated three recent measurement innovations to monitor wheelchair athletes. The data collected is used in feedback tools for athletes, trainers, and coaches, and enables them to optimize training and wheelchair settings for better performance.

In the domain of Energy, the paper "Improving Power System Resilience with Enhanced Monitoring, Control, and Protection Algorithms", Veera Kumar *et al.* presents key applications of so-called synchrophasors for enhancing the reliability and resilience of future power systems, covering real-time data-driven disturbance detection and blackout prevention.

In the paper "RATE-Analytics: Next Generation Predictive Analytics for Data-Driven Banking and Insurance", Collaris *et al.* present their collaboration between two universities, a bank and an insurance company. The overall goal of this project was to develop foundations and techniques for next generation big data analytics in banking and insurance. The main challenge of existing approaches is the lack of reliability and trustworthiness: if experts do not trust a model or its predictions they are much less likely to use and rely on that model. Hence, the project proposes solutions that bring the human-in-the-loop, enabling the diagnostics and refinement of models, and support in decision making and justification.

The final paper of the book, "Digital Art Technical Sources for the Netherlands: Integration and Improvement of Sources on Glass for a Sustainable Future – Art DATIS" by Capurro *et al.*, provides an overview of their Art DATIS project, through which they explore methods for the automatic transcription and documentation of diverse archival materials, focusing on the archive of Dutch glass artist Sybren Valkema. By digitizing documents and making their content searchable, the project aims to uncover how traditional glass making knowledge and practices evolved during the twentieth century.

It has been a pleasure for us to work with outstanding authors who enthusiastically contributed to this volume. Their dedication and open-mindedness have made this project rewarding and successful. We would also like to express our gratitude to the the members of the program committee and the reviewers, as well as to the Dagstuhl Publishing team, in particular Michael Wagner, for their support.

Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, Ruben D. Vromans.
September 2024, Tilburg & The Hague, the Netherlands.

Program Committee & Reviewers

Inigo Bermejo
Maastricht University

Carlos Budde
University of Trento

Roberto Casaluce (reviewer)
Scuola Superiore Sant'Anna, Pisa

Sven Dupré
Utrecht University, Utrecht

Paola Grosso
University of Amsterdam

Boudewijn R. Haverkort (chair)
Tilburg University

Paul Havinga
University of Twente

Aldert de Jongste
ECP, The Hague

Eline Meijer
Leiden University Medical Center

Maaike Snelder
TU Delft

Ruben D. Vromans
Tilburg University

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Use-Inspired Research on Big Data and Applications in the Public-Private Research and Innovation Program Commit2Data

Boudewijn R. Haverkort
Tilburg School of Humanities and Digital Sciences, Tilburg University, The Netherlands

Aldert de Jongste
ECP, The Hague, The Netherlands

Pieter van Kuilenburg
ECP, The Hague, The Netherlands

Abstract

In this paper we give an overview of the public-private research and innovation program known as Commit2Data, which was executed throughout the years 2016 – 2024 in the Netherlands. We outline the set-up of the program, with special attention for its valorisation activities, and provide a future outlook.

2012 ACM Subject Classification Information systems; Applied computing

Keywords and phrases Big data, public-private partnership (PPP)

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.1

Acknowledgements Running the Commit2Data program has been quite an endeavor, but has also been very fulfilling. We are grateful for the generous financial support from NWO, the Ministry of Economic Affairs and Climate, and the Topsector ICT. Over the years, many dedicated staff members of these organizations, as well as from the Netherlands Organization for Applied Scientific Research (TNO) and ECP supported the Commit2Data program generously with their time and expertise. We are grateful for their efforts, which have been instrumental to the success of the Commit2Data program.

1 Introduction

Data science is an interdisciplinary academic field, focusing on the collection of data sets from systems or processes, extracting information from such data sets, and further developing such information to knowledge that can be applied in solving a wide range of problems [4]. Data science became recognized as an independent academic discipline around the 1990s. Since then, the amount of data generated on a daily basis has increased in a way that is hard to phantom, primarily through the emergence of the internet, in particular the world-wide web [1]. More and more people globally are using the internet on a daily (hourly, minute) basis, and the advent of the internet of things has resulted in additional data generation and data transfer that requires virtually no human interaction; this is nicely visualized in Figure 1. As a result, the total amount of (meta-) data generated around the world on a daily basis is currently estimated at several hundred million terabytes. Since 1998, the collecting, handling and processing of such large amounts of data is often referred to as "big data" [2, 7].

However staggering the amount of data available, data in itself is of little economical and societal value. Only when data is analysed, turned into information, then into knowledge and finally applied to specific problems, it becomes valuable in a wide range of application domains. In order to reach this value, expertise from the application field has to be brought together with expertise on big data handling and big data analytics. It is exactly this

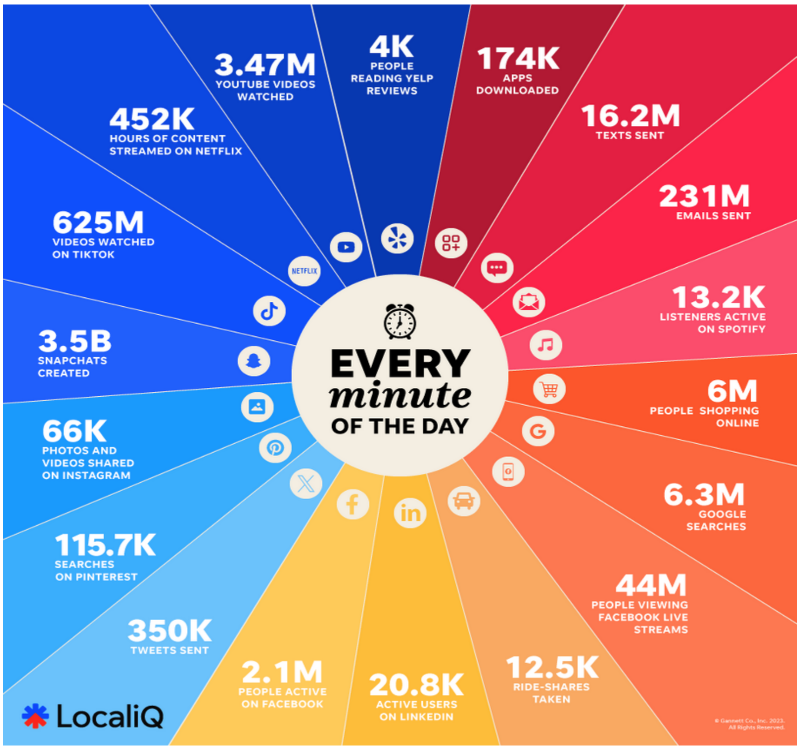


Figure 1 The amount of data generated per minute on the internet in 2024. Figure from `https://localiq.com/`.

combination that was put at the core of the Commit2Data public-private research and innovation program that was set-up throughput 2015 in the Netherlands [8], and that started in 2016.

This paper is further organized as follows. We will present the set-up of the Commit2Data program in more detail in Section 2. Subsequently, in Section 3, we will specifically address the valorisation activities we developed in the program. We conclude the paper with an outlook in Section 4. The papers following this paper will address a selection of projects from the Commit2Data program, across various application domains, in more detail.

2 The structure of the Commit2Data program

The Commit2Data program has been set-up in a joint effort of the Ministry of Economic Affairs and Climate (EZK), the Netherlands Organization for Scientific Research (NWO) and the Dutch national applied research and technology organization (TNO) in 2016. At that time, the expected impact of big data handling and big data analytics for all sectors, not just the ICT sector itself, was deemed enormous. It was also clear that many sectors were taking steps towards further datafication and digitalisation, yet many sectors were also struggling with this transformation. The aim of the Commit2Data program was in that sense twofold: to stimulate fundamental research in big data handling and analytics, as well as to stimulate the uptake of such advanced techniques in a variety of sectors, in particular

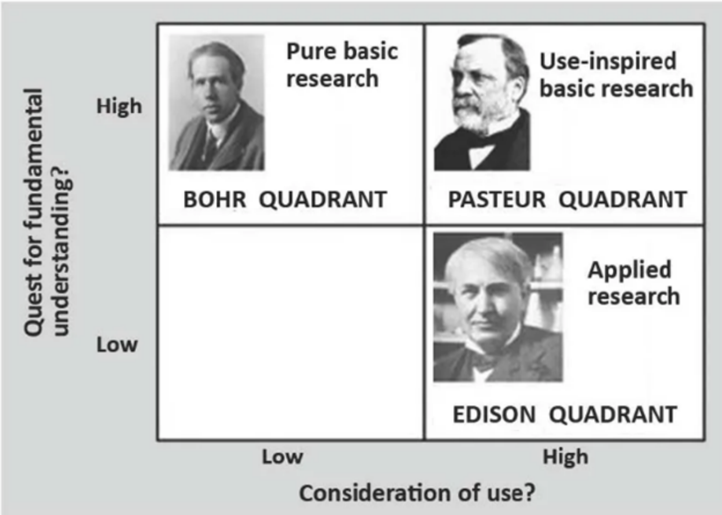


Figure 2 The four quadrants as described by Stokes; Commit2Data focuses on "use-inspired basic research", also known as the Pasteur quadrant.

the so-called Dutch topsectors.[1] By connecting the key enabling technology "big data" to the well-organized topsectors, a win-win situation was created: big data scientists, primarily coming from (applied) computer science departments at Dutch universities could connect to key applicants (often private parties) of such techniques, leading to fruitful cross-fertilization. In doing so, in the terminology of Stokes [6], so-called "use-inspired research" was performed, cf. Figure 2: basic research, yet inspired by concrete use-cases.

The Commit2Data program aimed to bring together (i) excellent data science, with (ii) specific application domain knowledge and inspiration. This is essential because the field of big data is delineated in two dimensions. The first dimension considers data properties and objectives, such as volume, heterogeneity and quality of data. Each of these comes with its own body of knowledge and scientific challenges. Together they define the science of data, data stewardship and data technology in any big data application. The second dimension is the application domain and context in which big data is to be used: contextual information is needed to understand the specific properties and limitations of the data and its intended use. Applications may be very different because of dissimilarities between application domains, yet the underlying subset of data science challenges is often very similar indeed. By addressing multiple application areas, as done in the Commit2data program, cross-fertilization and mutual gains can be obtained.

Figure 3 presents the overall structure of the program. Next to the initial two horizontal subprograms, each focusing on key challenges in the area of big data, without any specific application context in mind, most projects were developed in other subprograms (verticals), organized per sector. Per subprogram (in Figure 3: for every horizontal and for every

[1] The so-called topsectors are economic sectors with high activity and value creation in the Netherlands, responsible for a large share of the Dutch gross national product; see https://www.topsectoren.nl/ (only in Dutch).

vertical), open calls for research projects were issued. These calls were executed by NWO (the Dutch organisation for scientific research funding), however, the scope of the call was developed by dedicated working groups involving computer scientists as well as scientists from the respective application context. The subsequent call execution, including the selection of projects by NWO through independent reviews and independent selection panel making the final decisions, ensured high standards of scientific integrity. In the figure, the number of executed projects per sector is indicated; a project typically runs for 4–5 years, involves multiple universities and public and private partners, and gives financial room for the appointment of multiple PhD candidates and/or postdocs. The chairperson of the Commit2data program, Boudewijn Haverkort, was not involved in any of the selection procedures, nor did he participate in any of the projects himself, in order to maintain his independent role.

Before every call deadline, open matchmaking events were organised, to bring interested scientists together with scientists and engineers from the application fields. The funding received in each accepted project, was to be used for the appointment of PhD candidates or postdocs at the participating universities. Next to funding from NWO, participating public or private parties from the application domains, had to provide in-cash and in-kind co-funding. The amount of co-funding depended on the sector involved; some sectors are better equipped for high co-funding than others. The overall program budget amounted to 61M€, out of which 18M€ came from public and private partners, hence, we almost attained 30% overall co-funding. In this way, true public-private collaboration was created. This co-funding has multiple effects: for one, there is some influence of the external parties on the topics addressed in the project, however, the mere fact that these parties do really invest, makes their commitment to the project larger, e.g., in making data sets available, giving access to technical experts, etc. For the researchers, the possibility to work with real data, on industrial size cases, also puts their methods to the test, much more than in a smaller scale laboratory setting. Private parties involved in new technologies, like big data, also appreciate their interaction with universities, in order to keep abreast of developments and to position themselves as interesting employer. Finally, the involvement of the private parties from the outset of the project definition, makes that the research projects have high application potential; this has also been witnessed in the valorisation program, as will be discussed in Section 3.

Next to the organisation of research and valorisation, considerable effort was put on dissemination at program level. The aim has been to make the research conducted in the Commit2Data program accessible to a wider audience than just the consortium partners and readers of scientific articles. The dissemination output of the program encompasses a website[2] with extensive project updates, interviews with project leaders and researchers, various publications in trade journals, from healthcare to logistics to agriculture, and several animations that positions a project in a societal context and illustrate the practical societal value of academic research for a broader public.

Two parents: success factor and challenge. As stated, a key objective of the program was to organize use-inspired research. To achieve this, consistent efforts were made to connect data science knowledge and expertise, with in-depth knowledge of the application areas. To accomplish this, for most of the open calls, an exploration was conducted by a representative from the application domain, roughly addressing the question “What are the most significant

[2] The Commit2Data projects can all be accessed at `https://commit2data.nl/en/`.

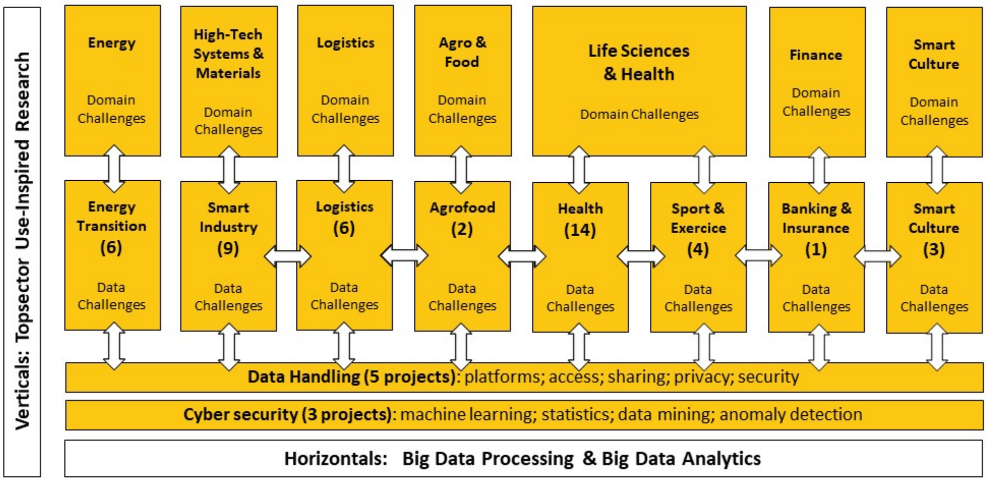


Figure 3 Overall structure of the Commit2Data program, with each block representing a number of thematically connected projects.

data-related questions within this domain?" Based on the outcome of such an exploration, we brought together sector-specific funding (e.g., from the topsector Energy, and the topsector Logistics) and Commit2Data funding. Together, they collaboratively drafted a call text, brought it to the attention of relevant parties, and organized a matchmaking event.

Whereas the two-parent approach has been instrumental in achieving the program's focus on true applications, it also has had a drawback. Many of the projects and many of the researchers in the projects, identify themselves as sector-specific projects (or researchers). They are primarily focused on healthcare or logistics (to name two examples), because that is where they feel their main network and interests lie. A resulting disadvantage of this is the perceived lack of cohesion within the Commit2Data research and innovation program as a whole. Originally, the idea was that knowledge exchange between sectors could also be organized, meaning that some sectors could benefit from results achieved in other sectors. After all, the idea is that a data analytics solution in healthcare might also work well for a similar problem in smart industry. However, the implementation of this cross-fertilization turned out to be challenging in practice. Project members felt their connection along the sector-specific line rather than along the data analytics line, and knowledge sharing outside the sectoral ecosystem turned out to be difficult to organize. As coordinators of the overall program, we would have liked to see this differently, and we still think that opportunities are missed, however, we did not manage to change this throughout the program. To conclude this section, Figure 4 visualizes some key indicators of the overall program.

3 Valorisation activities

Knowledge valorization is the process of creating value from knowledge, by making knowledge suitable and/or available for economic and/or societal use, by translating it into easy-to-use products, services, processes and new business [9]. Although the valorisation of research results was one of the program's objectives from the outset, no specific budgets were initially allocated for it. We therefore developed valorisation activities as a separate component from 2019 onwards, with separate funding. That was around the time that the initial projects started to deliver results suitable for valorization.

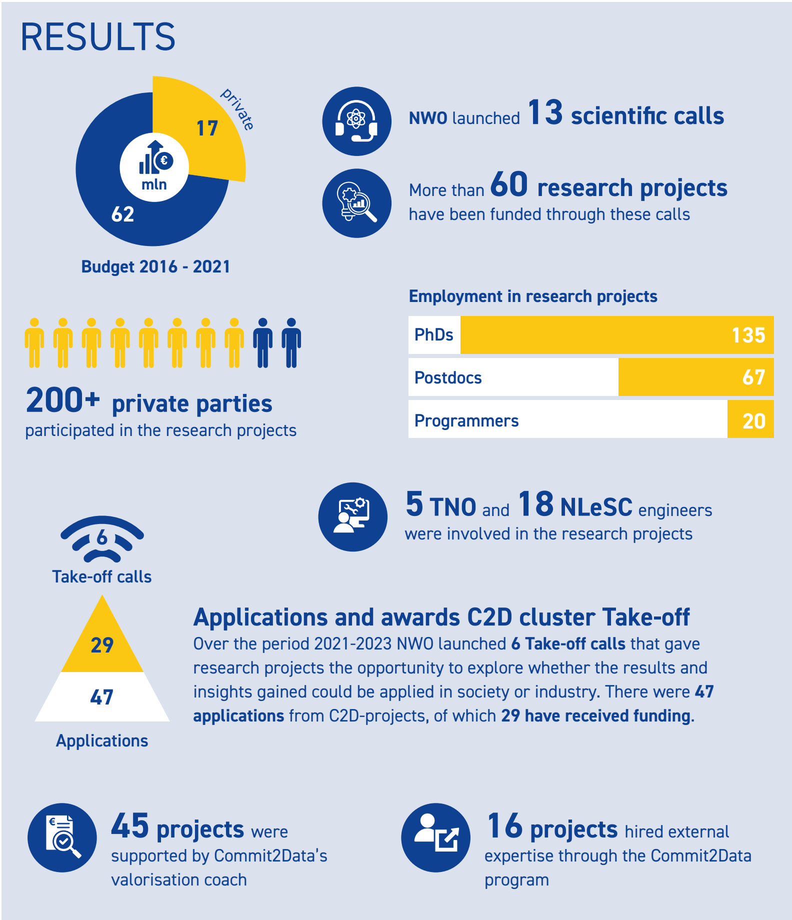


Figure 4 Visual of the overall output of the Commit2data program.

The development of the valorisation program raised several questions, e.g., about target groups, instruments and financing. For example, whom should we support, just the consortia as a whole or also individual project participants, like researchers, private companies or other societal actors? This in turn led to questions about the financing structure; none of the existing instruments in the Netherlands catered to all these different groups. Finally, a robust valorisation program could be established, building on the NWO Take-Off program, but the late start did have some drawbacks. Due to the late start, project consortia could only be informed and involved at a late stage of the projec execution. This meant that the consortia could only incorporate new possibilities into their approach in this later stage, and plan their staffing accordingly. Researchers are often committed for four years only; with full prior knowledge of the possibilities of the valorisation program, they can better plan ahead. A lesson learned from the Commit2Data program is therefore to have a clear understanding of what the follow-up to the research phase will be, already at the start of a program or call, so that project consortia can factor the use of such valorization options into their approach.

Although funding was pledged in various stages, all partners agreed on a valorisation plan by the end of 2019, which received an extra financial impulse from the ministry of Economic Affairs and Climate (close to 2 M€). The original plan included a variety of instruments, from raising awareness and influencing perceptions to customized support for projects or individuals facing specific challenges. The more generic instruments, such as valorisation workshops, were found to have limited reach within the project consortia, and by the year 2000, we transitioned fully to a customized approach. The central figure in this approach became the so-called valorisation coach. The valorisation coach, which we hired specifically for this purpose, approached project groups, starting with those projects that would come to an end first and were therefore closest to the potential useful results. Projects that started later were approached later. This approach was also partly based on the availability of capacity for support.

The valorisation coach had various tools for support, including his own experience and network. In practice, he often assists researchers, but also startups or consortia in sharpening their valorisable results and making valorisation challenges more manageable. The plan included additional financing so that topic-experts could be hired for specific questions to help and to advise projects or even develop proof of concept implementations. On several occasions, an experienced entrepreneur was sought and found who could coach a startup with the help of our funding.

Substantial effort was put into setting up a valorisation granting scheme, through which researchers could apply for up to 40 K€ to develop a research idea further into a product or first prototype for demonstration purposes, as a first step to further societal or commercial exploitation. A special call for the purpose was issued twice per year in the period 2022–2024 (in total 6 calls) via the NWO Take-Off program. From within the Commit2data program, 46 applications were submitted for such support, of which 27 were granted. The Commit2Data valorisation coach did play an active role in encouraging and supporting projects to submit good proposals to these calls. As a result, not only is the number of applications exceptionally large, also the success rate is very high, which we explain by the fact that our projects have been set-up to be use-inspired, and by the active involvement of our valorisation coach, respectively. Figure 4 also summarizes the key output of the valorization program.

4 Future outlook

The field of data science has undergone significant evolution, marked by the proliferation of big data and the increasing prominence of artificial intelligence (AI). In contemporary discourse, many of the techniques developed in the context of big data research, are now subsumed under the broader and more fashionable term "AI." This conflation, although technically inaccurate describes the current state of practice.[3] Many of the initiatives that started in the Commit2Data program, have found their way to the Netherlands AI Coalition (NLAIC),[4] and their funding schemes, like AINed[5].

Even though the term "big data" is less fashionable these days, big data is poised to continue its impact across various sectors. Advances in data storage and processing technologies, such as cloud and edge computing, are expected to handle even larger and more complex datasets with greater efficiency. Further integration of big data with technologies like the internet-of-everything[6] and the wider availability of bandwidth for mobile communication

[3] In the interview [5], machine-learning pioneer Michael Jordan explains why today's artificial-intelligence systems are not actually intelligent; see also [3].
[4] See `https://nlaic.com/en/`; the NLAIC is a public-private partnership in which knowledge institutes, government agencies, societal organisations and commercial companies work together to accelerate AI developments in the Netherlands and connecting AI initiatives.
[5] See `http://ained.nl/en/about-ained/`.
[6] See `https://ioe.org/`.

will enable real-time data collection and analysis on an increasing scale. As the interweaving of big data technologies with our everyday lives grows, so does the need for proper ethical considerations. To safeguard data privacy and security, innovative and robust frameworks will have to be developed to ensure responsible data governance. As these trends unfold, the distinction between big data and AI will likely blur further, fostering innovative applications and insights.

The Commit2Data program played a significant role in advancing methodologies and applications with big data. The approach developed through Commit2Data are now more commonplace in a range of funding calls by the Netherlands Organisation for Scientific Research (NWO). These calls reflect a broader integration of data science principles, emphasizing the importance of data-driven decision-making and the application of sophisticated analytical techniques across diverse scientific fields. This progression underscores the dynamic nature of data science and its foundational impact on contemporary research paradigms.

References

1 Tim Berners-Lee, Robert Cailliau, cois Groff Jean-Fran and Bernd Pollermann. World-Wide Web: The Information Universe. *Internet Research*, 2(1):52–58, 1992. doi:10.1108/eb047254.

2 Francis X. Diebold. On the origin(s) and development of the term "Big Data". *Penn Institute for Economic Research Working Paper*, 12(37), 2012. doi:10.2139/ssrn.2152421.

3 Michael I. Jordan. Artificial Intelligence – The Revolution Hasn't Happened Yet. *Harvard Data Science Review*, 1(1), July 2019. https://hdsr.mitpress.mit.edu/pub/wot7mkc1. doi:10.1162/99608f92.f06c6e61.

4 Chantal D. Larose and Daniel T. Larose. *Data Science Using Python and R.* Wiley, 2019.

5 Kathy Pretz. Stop calling everything AI. *IEEE Spectrum*, March 2018.

6 Donald E. Stokes. *Pasteur's Quadrant – Basic Science and Technological Innovation.* Brookings Institution Press, 1997.

7 Michael Stonebraker, David Blei, Daphne Koller, and Vipin Kumar. ACM Panels in Print: Big Data. *Communications of the ACM*, 60(6):24–25, 2017.

8 Topsector ICT. Commit2data white paper: Proposal for a national public-private research and innovation program on data science, stewardship and technology across top sectors. Technical report, Topsector ICT, the Netherlands, 2015. URL: https://ecp.nl/publicatie/commit2data-white-paper/.

9 Leonie van Drooge and Stefan de Jong. Valorisatie: onderzoekers doenn al veel meer dan ze denken. Technical report, Rathenau Instituut, Den Haag, 2015. URL: https://www.rathenau.nl/nl/kennis-voor-transities/valorisatie-onderzoekers-doen-al-veel-meer-dan-ze-denken.

Building a Digital Health Twin for Personalized Intervention: The EPI Project

Jamila Alsayed Kassem
MNS, University of Amsterdam, The Netherlands

Corinne Allaart
Vrije Universiteit Amsterdam, The Netherlands
St. Antonius Ziekenhuis, The Netherlands

Saba Amiri
MNS, University of Amsterdam, The Netherlands

Milen Kebede
CCI, University of Amsterdam, The Netherlands

Tim Müller
CCI, University of Amsterdam, The Netherlands

Rosanne Turner
CWI, Amsterdam, The Netherlands
UMC Utrecht, The Netherlands

Adam Belloum
MNS, University of Amsterdam, The Netherlands

L. Thomas van Binsbergen
CCI, University of Amsterdam, The Netherlands

Peter Grunwald
CWI, Amsterdam, The Netherlands
Leiden University, The Netherlands

Aart van Halteren
Vrije Universiteit Amsterdam, The Netherlands
Philips Research, Eindhoven, The Netherlands

Paola Grosso
MNS, University of Amsterdam, The Netherlands

Cees de Laat
CCI, University of Amsterdam, The Netherlands

Sander Klous
University of Amsterdam, The Netherlands
KPMG, Netherlands

Abstract

The Enabling Personalized Interventions (EPI) project, part of the COMMIT2DATA top sector initiative, brings together research on data science, software-defined network infrastructure, and secure and trustworthy data sharing, executed within the healthcare domain. The project applies the digital twin paradigm, in which data science-driven algorithms monitor and perform functions on a digital counterpart of a real-world entity, to enable proactive responses based on predicted outcomes. The EPI project applies this paradigm in the healthcare context by developing and testing applications that can act as personalized digital health twins for self/-joint management.

The EPI project addresses several challenges to digital twin applications in the healthcare domain, such as: 1) strict health data sharing policies often lead to data being locked in silos, 2) legal, policy and privacy requirements make data processing increasingly more complex, and 3) significant limitations on infrastructure resources may apply.

In this paper, we report on the use cases the EPI used as the basis to develop possible solutions to these challenges. In particular, we describe algorithms and tools for algorithmic real-time response and analysis of distributed data at scale. We discuss the automatic enforcement of legal interpretations and data-sharing conditions as executable policies. Finally, we investigate infrastructural challenges by implementing and experimenting with the EPI Framework - consisting of a distributed analysis infrastructure and BRANE for orchestrating multi-site applications. We conclude by describing our Proof of Concept (PoC) and showing its application to one of the EPI use cases.

2012 ACM Subject Classification Information systems → Data exchange

Keywords and phrases Healthcare, Data Sharing, Personalised Medicine, Real-time Data Analysis, Digital Health Twin, Data Policies

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.2

Funding The EPI project is funded by the Dutch Science Foundation in the Commit2Data program (grant no: 628.011.028).

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 2; pp. 2:1–2:18
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Recent advances in healthcare are largely due to improved technologies that allow for early diagnosis [10]. Data analytic technologies have strong potential to help solve complex problems in many industries: knowledge is power – and in healthcare, that holds true in particular. Yet, for an industry that is under financial stress, the increasing complexity of disease and comorbidity, data-centric advances in health have been burdened by capacity constraints – why has data analysis not been healthcare’s saviour? A number of challenges have inhibited this. For example, data is not accessible and remains (predominantly) in silos; data is not widely analysed to derive meaningful clinical insights; insights are not accessible as actionable information for healthcare providers or patients to self/joint manage the patient’s condition. The project described in this publication - EPI - Enabling Personalized Intervention[1] - addresses these challenges with the ultimate objective of improving cost efficiency, quality, and outcomes of care, while ensuring patient and public health data and results are processed safely and with respect for the digital (privacy) rights of patients.

The EPI project vision is structured around the concept of the *digital health twin* (DHT) as defined by [5]. “Digital Twins stand for a specific engineering paradigm, where individual physical artefacts are paired with digital models that dynamically reflect the status of those artefacts.” This leads to the hypothesis that with a DHT: “one would be in the possession of very detailed bio-physical and lifestyle information of a person over time. This perspective redefines the concept of “normality” or “health” as a set of patterns that are regular for a particular individual, against the backdrop patterns observed in the population.” The development of personalized digital health twin algorithms will inherently benefit from the ability to collect as much relevant data as possible in order to better cluster patient groups, and cater diagnosis treatment and outcome prediction according to the patients’ genetic makeup and medical history. Given the emphasis on the individual and on the dynamic nature of a DHT, our main research question is if the existence of a DHT indeed enables instant (real-time), effective, personalized guidance to prevent health related incidents and/or helps improve intervention effectiveness.

The paper starts by going over the project organization in Section 2, then we introduce the DHT use cases investigated in Section 3. We address the challenges of building a real-time algorithmic response to enable collecting and analysing patients’ data within a DHT use case in Section 4. On the other hand, data is often distributed and hosted across healthcare domains, and we discuss the methods to analyse data at scale utilizing distributed learning algorithms and privacy preserving machine learning in Section 5. In Section 6, we introduce automated policy interpretation, management, and enforcement to enable data usage abiding to set policies and defined purposes. On a lower level, we address infrastructural heterogeneity and security-based challenges by formalizing the data-sharing framework built with BRANE in Section 7, and the security service orchestrator in Section 8. Finally, we show the Proof of Concept of the framework running the use cases in practice in Section 9.

2 Project organization: partners roles and partners interactions

The EPI project was funded by the Commit2Data Data2Person call from NWO, the Dutch national funding agency. The project partners started their work in 2019 and are now in the concluding phase of their collaboration.

[1] `https://enablingpersonalizedinterventions.nl/`

Already in the preparation phase of the project, it was clear to all involved that the only chance to develop a working solution was to create a cooperation of an interdisciplinary team of medical specialists, data scientists, ICT infrastructure experts as well as experts in artificial intelligence and law. In accordance with that, the EPI project brings together academic and research partners, healthcare providers and the private sector to tackle the challenges related to enabling personalized intervention.

The University of Amsterdam (UvA), the Centrum Wiskunde and Informatica (CWI) and the Vrije University Amsterdam (VUA) provide their scientific expertise and employ PhD researchers to work closely with the other project partners. The three hospitals within the consortium, St. Antonius, Princess Maxima Center, University Medical Center Utrecht (UMCU) provide three representative health-related case studies. The use cases all have their own requirements for the operational analysis environment. It is obviously not scalable nor desirable to have to manage yet another environment for each use-case: the infrastructure developed as part of EPI solves that problem. To accomplish this the medical institutions have put their domain experts in contact with the researcher at the universities and created close collaborations on the individual use cases. SURF, also involved in the project, provides the infrastructural components; it fulfils the role of platform manager, established to manage the operation of the platform and to monitor and enforce that its participants behave as expected. Finally, KPMG and Philips as commercial partners in the project contribute insightful direction towards long term sustainability of the EPI results. During the whole duration of the project, all partners have maintained very close interactions, even in periods of Covid lockdowns.

As the project is now in its concluding phase the partners are actively working for the results to see broader adoption, beyond the time span of the project. It is clear that EPI offers models, tools, and software that are also broadly applicable, not just to the medical field. A first and concrete outcome of these efforts is that the collaboration of EPI with AMdEX[2]. AMdEX is an innovation field lab initiated by AMS-IX, SURF, UvA, DEXES and Amsterdam Economic Board and co-funded by the European Regional Development Fund. The EPI partners have already acquired funding for integration of the EPI result in AMdEX, and are looking for further commercialization possibilities.

3 Use Cases

3.1 Personalized predictions for stroke rehabilitation (St. Antonius Hospital)

The goal of this Use Case is to provide personalized outcome predictions for the rehabilitation of patients that suffered from a cerebrovascular accident (CVA), utilizing the data of these patients that is distributed among different care institutions, in a privacy preserving manner.

Outcome predictions based on the complete care path. After the acute care for a CVA in the hospital, patients often receive therapy in the form of in- or outpatient rehabilitation. As CVA patients are a diverse group of people in different conditions and different stages of life, the rehabilitation treatments and the long-term health outcomes after rehabilitation vary [28]. While both patients and healthcare professionals express the importance of personal

[2] This work is partially funded through the AMdEX Fieldlab project supported by Kansen Voor West EFRO (KVW00309) and the province of Noord-Holland. `https://amdex.eu`

prognoses, little is known about these outcomes at discharge from the hospital. To provide personalized outcome predictions for these patients, and create an overview of the chain of care, this project aims to develop a DHT that includes their entire care path. However, this is complicated due to the involvement of multiple healthcare institutions in these paths, including hospitals, rehabilitation clinics and nursing homes. The data of each patient is spread out among these institutions, and organizational, compatibility and privacy issues will often surface. We need to create personalized outcome predictions while only sharing information in a privacy-preserving manner. This project aims to achieve this by developing a distributed method to create prediction models while retaining the privacy of the data of the participating clinics.

Clinical interface to assist with shared decision-making. Within health care, there has been a movement to shift towards shared decision-making (SDM) [12], where the health care professionals create a care plan not only based on their own clinical expertise but also actively taking into account the personal needs and want of each individual patient. For patients who have experienced a CVA, a large, heterogeneous patient group, the rehabilitation in post-acute care is a situation where SDM can be beneficial [27] To aid the SDM process, we aim to develop a clinical interface that provides personalized insights into the prognosis of the patients, dependent on the different care paths. The basis of this interface will be developed using distributed prediction models. These models will depict the predicted outcome of different individual care paths to aid clinicians during the SDM process.

3.2 Personalized recommendations in Psychiatry

The goal is to provide personal and explainable treatment recommendations with the aim of accelerating the process of finding the right treatment for individual psychiatry patients.

Determining the information that the clinical interface should display. We aim to develop a clinical interface where, based on their DHT at the psychiatry department, patients and their clinicians receive insightful treatment recommendations and predictions. What the nature of these recommendations should be in order for them to be the most useful to patients and clinicians is not entirely clear. In psychiatry, there is currently no consensus on the best ways to measure and predict treatment effects, especially not for large, heterogeneous groups of patients with multi-morbidity. Currently, in clinical trials, hundreds of different kinds of symptom rating scales (which are not suitable for heterogeneous groups of patients) and quality of life rating scales are used to assess treatment effects, as well as measures on autonomy, social participation and the occurrence of side effects [35]. To work toward a solution, interviews with patients and clinicians at the participating mental health hubs were carried out to determine which of these different treatment outcomes would be most useful to predict and present to patients and clinicians. The feasibility of extracting the proposed outcome measures from the digital health twin was explored; could the information be extracted from the mix of structured and unstructured data available in the digital health twin? And could this be done in real time, i.e., when new patient outcome data becomes available, for example when a patient reports to the clinician that they experience side effects, would this information then be transferred to the digital health twin, and could the recommender system use this to adapt the predictions? Results of this study are found in [35].

Developing personalized recommender systems for psychiatry patients. Psychiatry syndromes are heterogeneous, and side effects play a major role in treatment adherence and outcome. Hence, personalizing the predicted outcomes of the recommender algorithm could vastly improve the time to find the right treatment. One patient might be interested in experiencing as few side effects as possible, while another might be mainly interested in going back to work as soon as possible. In a third patient, a clinician might recognize an overlap of two syndromes, where they know the interaction between two symptoms plays a crucial role in the persistence of the disease, and they aim to target these specific symptoms early during treatment. To enable this, we aim to develop and train a recommender algorithm that predicts the best treatment for a given patient, for either one outcome or a mixture of outcome measures. In the clinical user interface, the clinician and patient will then be able to decide on this weighted mixture together, before receiving the recommendations.

Sharing recommender systems between mental health hubs. After a recommender system has been trained and validated it will be shared with other mental health hubs, with two aims. Firstly, before medical recommender systems can be used in practice, their generalizability needs to be demonstrated through validation in an independent cohort of patients. Secondly, after validation, the algorithm should be shared such that it can be used to enrich the digital health twins in other mental health hubs.

To enable the exchange of the algorithm, a secure infrastructure should be employed, where all the mental health hubs' policies are enforced, e.g. it is controlled who has access to which parts of the recommender system. For example, it should be controlled that data from individual patients cannot be retrieved through (requests to) the shared algorithm (see section 7). The sharing platform should also integrate the different informed consent systems from the mental health hubs. Another major challenge, not currently addressed in this project, is the potential difference in the structure of the digital health twins employed by each hub. When the recommender system takes structured data as in- or output, this will cause problems. For example, in one hub, previous medication of a patient might be stored in a structured manner by the pharmacist, whereas in another hub, previous medication might be written down by the clinician in a free text field in the electronic health records. When our system is then built in the first hub, it will not 'know' where to find the right information in the second. Preprocessing the data of the digital health twins in each hub is a process that might have to be automated for future real-time applications, possibly through introducing a meta-layer around the recommender algorithm.

3.3 Princess Maxima Center

The EPI project's goal for this use-case is to enable compliant data-sharing by automating privacy policies from privacy legislation and contractual agreements on data sharing in order to facilitate stakeholder collaboration in discovering new treatments and prognosis factors for the Diffuse Intrinsic Pontine Gliomas (DIPG) disease, a rare and deadly childhood malignancy[19]

A Data sharing agreement specification for compliance and interoperability of access to data. Despite the fact that there have been almost 40 years of single-centre and non-randomized trials, there has not been significant improvement in survival rates of DIPG patients [16]. Consequently, there is an urgent need to share data for international collaborations to discover new treatments, prognosis factors, and other significant achievements. For this reason, the European Society for Paediatric Oncology (SIOPE) Brain Tumour

Group established a DIPG registry and image repository to collect data of DIPG patients [3]. However, strict privacy laws and high non-compliance fees are obstacles to seamless data sharing among stakeholders. The automation and enforcement of such privacy policies can possibly reduce the amount of non-compliance caused by, either intentional or non-intentional, violation of such rules. Unfortunately, current data sharing infrastructures and policy reasoning mechanisms, to some extent, fail to provide the necessary languages and methods to enforce privacy rules. These challenges become more prevalent when sources of legal norms are considered, such as Data Sharing Agreements (DSAs) and privacy regulations (e.g., the GDPR).

In this work, a domain specific language, eFLINT is used to specify data sharing policies expressed in DSAs [40]. The eFLINT language is developed to formalize legal norms, as such we found the language suitable to specify data sharing rules, conditions and duties specified in DSAs and privacy regulations. The extensions made to the eFLINT language allow us to connect higher level and abstract privacy policies to lower level system policies, such as read and write access rights [39].

A purpose-based access control mechanism. Current traditional access control models lack the necessary mechanism to specify and enforce data access and usage policies specified in privacy legislation and contractual agreements. In this work, we propose an access control model that simplifies the process of granting researchers access to datasets based on research purposes. The main component of this model is the specification of higher level abstract purposes as system level action or sequences of actions which are the basis for purpose based access control policy.

Develop a user-interface. The current user interface for the DIPG use-case was developed by the AMdEX fieldlab project. The prototype can be used by all stakeholders, a researcher and the DIPG executive committee to submit project proposals and approve project proposals respectively. Additionally, our main objective, to allow researcher access to data based on the approval of a project proposal, has been implemented on the prototype.

4 Algorithmic real-time response

A DHT involves collecting and analysing patient data in real-time. Predictions of treatment effects for individual patients can be improved in real-time as the amount of available information keeps increasing. For these predictions to be useful, adequate information about their uncertainty and rationale should be provided to patients and clinicians. Unfortunately, there are currently many challenges with the way statistical learning and hypothesis testing techniques can be applied in healthcare, and merging real-time response with an adequate confidence measure is one of them. When we peek continuously at our results throughout an experiment ("real-time" analysis), biased and misleading results are retrieved with classical hypothesis testing methods [43]. Similar problems arise when we want to combine insights collected over multiple healthcare facilities, or when we want to adapt our experiments post-hoc [34, 14].

A new framework of hypothesis testing and estimation that can deal with these challenges especially well is called "safe anytime-valid inference" (SAVI) [32]. This framework allows for dynamically analysing studies as their designs may change during data collection and analysis while avoiding the previously mentioned misleading results. SAVI tests may also be extended to estimation through anytime-valid confidence intervals: confidence intervals that provide robustness, at any time during a study [17]. Further, multiple SAVI tests can easily be combined into one big test, for example when researchers are not willing to share an entire dataset, but only summary statistics. SAVI tests also allow for increasing test power when datasets collected from several institutions are heterogeneous. They can, for example, handle switching of outcome measures or predictors, when the data sources available for research change dynamically over time. We are developing models that can capture confidence of intervention outcomes for (small groups of) patients, with guarantees on the error bounds of the model [37, 36]. The resulting 'patient-tailored' advice can be presented through an application as part of the clinical user interface, where input and outcome measures can be adapted dynamically by all main parties involved: the patient, clinician and researcher. The SAVI test or confidence results in such an application can even be combined with SAVI results from other healthcare centres, not affecting the reliability of the estimates.

So far we have discussed inferential models for studies that are ongoing, in real-time. We also consider historical, observational data to build a model that can be adapted later. As data infrastructure and data sharing possibilities keep improving in healthcare, an abundance of this kind of data will become available as well. However, with observational data, the discovery of (treatment) effects might be masked by confounding factors or bias due to (unknown) fallacies in study design. These issues go beyond straightforward statistical significance and make it necessary for learning models to provide a form of explainability, i.e. the rationale for a model's decision needs to be transparent to clinicians, healthcare researchers and patients [11]. To this end, we explore the application of more complex methods that also provide strong explainability properties, such as constraint-based Bayesian networks [4]. Multiple predictors and outcome variables of interest can be included in these models, and the associations between them can be presented to the user in the form of a directed acyclic graph. With the resulting network, groups of or even individual patients that are more likely to respond well to certain treatment strategies can be highlighted.

In summary, for the algorithmic real-time response in the DHT, we strive to develop relatively simple and explainable mathematical models that provide robust recommendations and insights to aid decisions and expand knowledge on clinical diagnostic methods and therapies. The SAVI tests for inference and the network-based models incorporate prior knowledge based on expert knowledge and pre-training of the models, resulting in a low computational cost when using the algorithms in real-time. As both proposed models offer a lot of flexibility, new insights gained during real-time analysis can be used to update the models dynamically and change the hypotheses tested and the nature of the predictions made, while maintaining robustness and guarantees on estimation error bounds.

5 Analysing distributed data at scale

Besides the need for a real-time response, another challenge that a DHT has to address is scale. Specifically, for more advanced machine learning and deep learning methods to be accurate, large amounts of data are necessary for training and evaluation [44]. Yet, in health care, large collective sets of medical data are rarely available, because data is collected and stored in many different, independent institutions and strict privacy laws and regulations apply in regard to sharing this data to preserve the privacy of the dataset and patients participating in it. A solution for this is to take the computation to where the data is stored. In order to adapt to this setting, two main modifications need to be made. Firstly, the algorithms themselves need to be altered to be able to train themselves on distributed data. The form of data distribution influences the way the algorithm should be adapted. Second, we need to make the machine learning mechanism itself inherently privacy preserving to ensure no information is leaked by simply exposing the predictions or other outputs of these models. These are the two topics we address in this section.

5.1 Distributed learning algorithms

Data can be distributed "horizontally" or "vertically" among parties. With horizontal partitions, each partition contains a number of instances, but has all features of this instance, as in section 3.3. There, the dataset includes patients from different hospitals in multiple countries, but for every patient, their individual data is only recorded in one hospital. This is the opposite in vertical partitions, where features of one instance are spread out over different parties, such as in the CVA use case where the data of a patient is spread out over the different healthcare institutions they visited. Distributed learning has focused on these challenges by developing methods that deal with this data parallelism, where collaborative models are trained on distributed data while the individual parties keep ownership of their local data. Most notable has been the development of federated learning [26], which is based on local parties creating their own models that, through iterative model parameter sharing, converge to one central model. The developed models for horizontal learning have a similar predictive performance as models trained on fully centralized data [21]. Distributed learning on vertically partitioned data has so far been less explored [21], for it comes with additional complexity: the feature set of the participating parties is different, so they cannot simply all train the same model and exchange parameters. Particularly, it is common in vertically partitioned data that only one party has the label that the model is trying to predict. Recently, several studies have developed methods to utilize vertically partitioned data for deep learning, but these distributed learning solutions still have issues in terms of predictive performance, privacy and efficiency [15, 6].

For this project, we adapted and evaluated Vertical Split Learning, where a neural network is distributed over the locations, similar to the data [6], for a set of medical and other use cases. We saw that depending on the use case, feature distribution and models used, predictive performance could vary greatly [2]. This shows it is essential to properly test vertically distributed learning techniques, and there is still a need for more robust vertically federated learning techniques. This requires not only the adaptation of distributed learning methods to retain predictive performance but also the implementation of appropriate privacy and security methods. As the goal of this project is to make DHTs broadly applicable, this includes meeting requirements for various forms of vertically partitioned data, as in the stroke rehabilitation use case. Therefore, we aim to further develop and adapt distributed learning methods that can use vertically partitioned data, while retaining robust predictive performance, as well as privacy and efficiency.

5.2 Privacy Preserving Machine Learning

In recent years, distributed machine learning methods have been gaining traction due to them being able to address either or both principal concerns in data analysis: privacy and scalability. On top of that, there has been a paradigm shift in large-scale, big-data computation towards distributed and cloud systems [20]. The increasing maturity of these methods in other domains makes employing them in a healthcare context - with its inherent need for privacy and big-data processing needs [31] - extremely compelling.

A whole new area of research has developed to study privacy aspects in machine learning under the name "Privacy Preserving Machine Learning (PPML)". PPML-based models usually employ methods based on Cryptography and/or Perturbation to preserve the privacy of the data and/or model. Cryptography-based approaches for PPML utilize cryptographic methods and protocols to encrypt data while perturbation-based methods work by transforming the data, usually by adding some noise to the training data, algorithm parameters or the algorithm output. There are other (less explored) methods falling outside these categories, e.g. Synthetic Data [45], private aggregation of Teacher Ensembles [30].

Although several PPML methods have proven to serve their purpose, each comes with its own caveats. For these methods to work, not only do we need to enable our machine learning models to work on encrypted, perturbed or otherwise-transformed data, we also need to ensure that the model trained on the garbled data has similar performance and generalizability as the same model trained on unprotected data. If it works at all, the extra compute overhead resulting from taking these measures is certainly a downside. Especially in healthcare, parallel to the crucial need for guaranteeing preservation of privacy, the performance of a model in terms of accuracy and its computational needs is a decisive factor for its adoption in the field.

Our aim is to not only provide privacy-preserving methods of machine learning suited for medical use-cases, but also to provide a framework for setting up a PPML pipeline tailored to the needs of our stakeholders. This framework will serve the purpose of preservation of privacy in distributed/federated machine learning scenarios. The distributed nature of the data brings up new challenges, e.g. bias and class imbalance, potential unreliability of learning parties, communication costs, all of which need to be addressed specifically for our target use-cases. This approach on the one hand enables algorithms introduced in Section 4 to run in a distributed privacy-preserving manner; On the other hand, it is integrated closely in a decentralized distributed scheme with the policy specification (Section 6), infrastructure (Section 7) and orchestration (Section 8) layers to ensure the PPML process adheres to governing rules and regulations and is correctly set up and deployed at stakeholders' site.

Adopting PPML in a DHT is in line with the remaining recommendations on mathematical modelling of [9]. PPML enables the simultaneous processing of disparate, distributed private datasets without creating privacy preservation concerns. Thus, it allows establishing private online repositories, including repositories containing sensitive data of individual patients, whilst guaranteeing data processing is privacy preserving independent of the analytical models built on top. A framework providing a common PPML foundation helps to seamlessly combine phenomenological and mechanistic models as it shifts the focus from the prevention of data leaks to the development of hybrid integration methods and strategies. Furthermore, it prevents privacy incidents when relations across scales are identified, potentially leading to homogenization and distribution strategies for the development of a theoretical framework for the analysis of scale separation.

6 Automated policy interpretation, management and enforcement

Current approaches to access to healthcare data for research purposes can be complicated due to strict privacy regulations and contractual agreements such as data sharing agreements. Privacy regulations and data sharing agreements are usually written in natural language, which can lead to ambiguous and inconsistent interpretations [1]. Resolving these challenges demands suitable policy specification languages to specify policies. The main objective of this work is to automate data access requirements from privacy regulations and data sharing agreements (DSAs) to reduce the complexity of data access for research purposes and enable the enforcement of different sources of legal norms via access control mechanisms.

6.1 Data sharing agreements and access control

The General Data Protection Regulation (GDPR) states the principles and rights that must be met for processing any personal data [7]. These principles are lawfulness, fairness and transparency; purpose limitation; data minimization; accuracy; storage limitation; security and accountability. The GDPR provides strict privacy norms, and there is no (access control) mechanism that enables the enforcement of all such norms. These challenges become more prevalent when several sources of legal norms (e.g., GDPR and DSAs) are simultaneously applicable, as is often the case with medical research. Additionally, when data is merged from several sources and different controllers are allowed to author their own policies that dictate for each source, it can cause policy conflicts. The gap between legal requirements and its technical realization is due to the lack of machine-readable representation of privacy policies.

The goal of this work is to formalize privacy policies and to enable their enforcement via (novel or existing) access control mechanisms. In the initial stage of our work, we analysed different policy specification languages to determine which of them met the privacy policy requirements from GDPR and DSA of the DIPG Registry use-case. The Open Digital Rights language (ODRL) [18] and the eXtensible Access Control Markup Language (XACML) [33] were concluded not to be sufficiently expressive. The analysis of ODRL is reported in [25]. Instead, we are applying the eFLINT language, originally developed within the SSPDDP project[3], to formalize GDPR and DSA articles relevant to the DIPG use case.

The eFLINT language is a domain specific language that was specifically designed to formalize legal norms [40]. The eFLINT language extensions enabled us to interconnect higher level privacy policies such as those expressed by the GDPR and DSAs with system level policies such as those specified by access control policies [39]. Additionally, we formalized the access request workflow of the DIPG registry use-case, a researcher submitting project proposals and approving proposals and the duties that follow as a result of these transitions. Finally, we demonstrate how actions described by the GDPR such as “collecting personal data” can be synchronized with actions from DSA such as “making data available” to the registry. This approach allows for the modular specification of norms by permitting re-use between specifications as well as defining alternative specifications.

6.2 Purpose Based Access control mechanism

In this work, we propose an access control model that simplifies the process of granting researchers access to datasets based on research purposes. Article 5(1(b)) states that personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes. Additionally, If personal data is collected for more than one purpose and these purposes are related, then the GDPR allows for processing to take place under an “overall purpose” under which a number of separate processing operations can take place. Therefore, the purpose limitation principle minimizes the risk that might arise when personal data is processed by confining the possibilities of its usage by limiting instances of lawful processing.

Our purpose based model is based on these requirements. The main component of this model is the specification of higher level abstract purposes as system level action or sequences of actions which are the basis for defining purpose based access control policy. We model action relationships based on the purpose specification requirements from the GDPR and

[3] The SSPDDP project (628.009.014) is part of the NWO program Big Data: Real-Time ICT for Logistics

express constraints over these relationships using the eFLINT language. The goal of this work is to design and implement an access control model that allows access to data based on the purpose of research plans.

7 BRANE - A heterogeneous, distributed research platform

Different DHT functionality requires different physical infrastructures to be deployed. On a software level, they have different hardware- and software requirements, such as dependencies on accelerators or specific software packages; and on an application level, they need different data flows and might impose different security requirements. Therefore, there is a need to dynamically map a logical infrastructure dictated by a DHT functionality onto a physical pool of resources. This pool of physical resources is often heterogeneous and dispersed as well. The exact composition varies but may consist of resources in the cloud, high-performant grid computers or even small, ARM-based devices that operate only locally. This variety introduces interoperability issues on the software level (different software stacks, transfer protocols or granted system privileges), hardware level (processor architectures, availability of accelerators, networking capabilities) and administration level (different costs, usage quotes or service-level agreements). Furthermore, heterogeneous trust relationships between resource providers may exist [29]. This restricts how data can flow between resources, or sometimes even if the provider of a resource is willing to do a task at all if this violates their own trust- or security policies. This issue becomes even harder when some of the policies involved are sensitive themselves and must thus be kept private; it may be, for example, that a dataset can only be processed by a resource if a patient has given consent to do so, where this consent itself already reveals sensitive information about the patient.

Traditionally, the mapping of a logical infrastructure on a physical pool of resources is done (semi-)manually. However, this approach does not scale in the case of DHTs, as both ends of the mapping are dynamic: new DHT functionality may be introduced or adapted, resources may be introduced or removed and the policies of existing resources may change. Hence, a mechanism is required to perform this mapping automatically and dynamically.

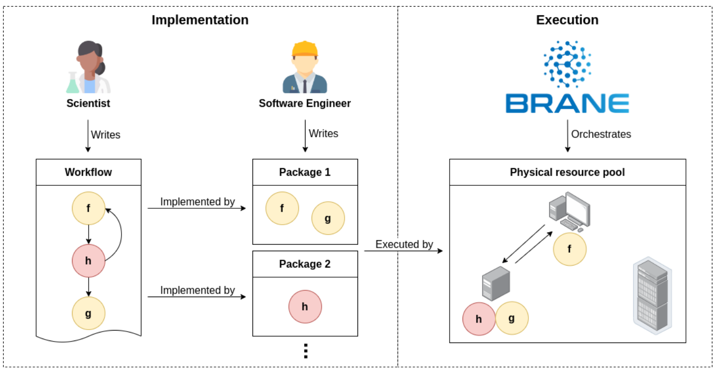


Figure 1 Conceptual role of BRANE. During implementation, a scientist writes a workflow that composes elementary functions (`f`, `g` and `h` in this example). These elementary functions are implemented by packages, which are written beforehand and made publicly available by software engineers. At runtime, the BRANE runtime understands the dependencies between the functions and can map them onto a physical pool of resources.

To solve the software and hardware interoperability issues, the BRANE Framework [38] is introduced as the mechanism to provide the logical infrastructure for physical resource pool mapping. It frames DHT functionality as *workflows*, which are compositions of elementary functions to form (data) pipelines that implement the desired behaviour. These elementary functions, in turn, are implemented by *packages*, which are containerized pieces of software such as data preprocessing steps or (parts of) algorithms that are executed on the resource pool. The runtime of the framework can then orchestrate these containers based on the requirements and dependencies laid-out by the workflow, providing the mapping discussed; this act is called *planning* a workflow. A visualization of this process is given in Figure 1.

To handle the administrative-, privacy- and trust-related interoperability issues, the BRANE framework is extended with a notion of *policy* [13]. These policies are expressed on a per-resource level, which effectively makes policies representative of a resource owner’s wishes within the system. By consulting with the policy both during the planning of a workflow, and during execution, the policies restrict what a resource is used for and what happens to the resource’s data, as well as express administrative limits or costs on the resource. A key feature of this notion of policy is that the policies are abstracted away behind a service, the *checker service* (Figure 2). Using an interface like this allows resource owners to make their policies arbitrarily complex since the system has no requirements on how they are implemented. For example, they can use languages like eFLINT [40] that can capture norms, allowing policies to express legal concepts like GDPR. Moreover, addressing the need for private policies, the interface allows every participant to hide as much of their policy as desired. This can range from “metadata” of a policy, such as the name of the person who gave consent, to the policy rules themselves; although the latter requires advanced reasoning about what policy information is leaked through the interface. This ability to keep policies private is what sets BRANE apart from similar workflow systems, like [41] or [8].

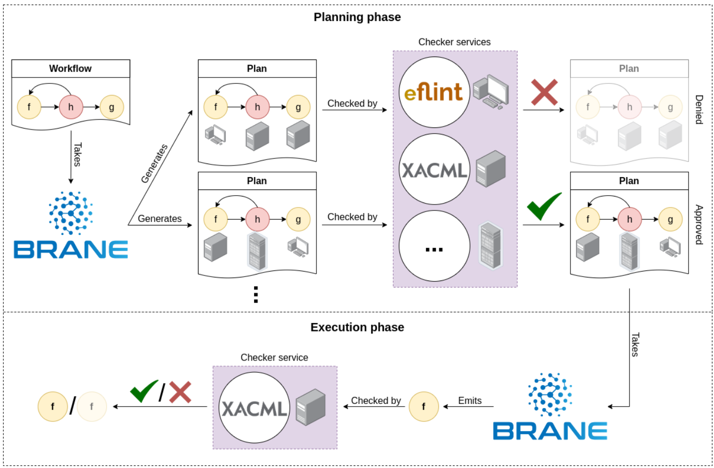


Figure 2 Policy enforcement in BRANE. There are two phases when executing a workflow: in the first, the *planning phase*, the runtime maps tasks in a workflow to resources that will execute the tasks. Here, checker services (and thus policies) act as filters to separate plans they would allow from plans they would deny, preventing the system from having to execute every possible plan. Then, during the *execution phase*, the runtime starts traversing the planned workflow and attempts to execute each task on the resource it was planned on. The checker of that specific resource examines the task again, to prevent any staleness introduced by the arbitrarily long delay between the planning-check and the execution of that specific task.

In summary, we present an automatic mapping of the logical infrastructure dictated by DHT functionality to a physical pool of resources using the BRANE infrastructure. It frames the functionality as workflows and packages, and complements that with high-level constraints on that mapping by introducing (potentially private) policies that are hosted on the resources themselves. To enforce lower-level policies related to (network) security, a more complex extension to BRANE is required and described in the following section.

8 Distributed Analysis Infrastructure

Currently, "working in silos" is still dominating the healthcare industry, but data-sharing between health domains is key on the road towards DHT. We aim to lay the foundation of distributed infrastructure and to utilize it in deploying the previously mentioned use-cases. In the context of this project, we define a data-sharing framework that allows parties in specific (research) networks to share patient information securely in an ad-hoc way in the course of treatment or a specific research question, as well as structurally for long-term studies, while maintaining control and tractability/suitability to data controllers and subjects at the source.

Developing a generic solution to comply with the dynamic security policies is difficult, mainly because the infrastructure between different healthcare domains is heterogeneous in terms of computing, networking and storage functionalities. Moreover, when various use cases have to be deployed (e.g. for running an ML algorithm, monitoring, sharing EHRs, etc.), different requirements or rule sets (policies) apply (e.g. privacy related) and have to be enforced to ensure proper control over the data flow/ usage. The infrastructure and security models need to support these different applications, therefore need to be adaptive to be able to support dynamic and often application-specific sets of requirements.

The main design properties we consider while setting up a secure, collaborative networked environment [23] between different EPI parties are: 1) the low level security and network policies formalizing and enforcement, 2) compatibility with higher-level policies, and dynamic workflow schemas, 3) evaluating data-sharing domains, and adaptively provision network service chains to maintain security and quality of service, 4) reacting to any policy change while flexibly trading and routing packets via the chain via proxies (different implementation presented in [24]). The proposed framework considers the intended data sharing use case, and the policies associated. We map the mandated network services to the defined enforcement primitives: *filter* traffic and/or *transform* traffic. The framework dynamically provisions these services by placing the functions on available N-PoPs (Network Points of Placements), assigning the service requests to the running function, and routing traffic along the function's chain to enforce a policy. We *optimize* provisioning decisions to maximize the quality of service based on the infrastructure state, the use case's requirements, and the CPU profiles of services [22]. The hyperparameters are tuned to prioritize resource utilization or latency in an effort to comply with the performance requirements. Three provisioning tools are used: a greedy heuristic approach, Deep Q-Learning (DQL), and a Heuristic-boosted DQL (HDQL).

We manage and orchestrate virtualized networked services on top of the existing client's infrastructure to increase the security level of the communication channels and enforce policies. A cloud-native network orchestrator is defined on top of a multi-node cluster mesh infrastructure for flexible and dynamic containerized function scheduling. The expected challenges include verification of rules with laws and international policies, integration with legacy systems, and acceptability of the framework by health institutions. The latter includes providing the health institutions with sufficient oversight and control over the sharing and usage of their data so that they are confident that this exchange is compliant and controllable. We address these challenges by working closely together with the responsible entities (IT departments, ethical boards, and data privacy officers) in the associated hospitals. The approach we will be taking is to formalize a logic to automate infrastructure setup

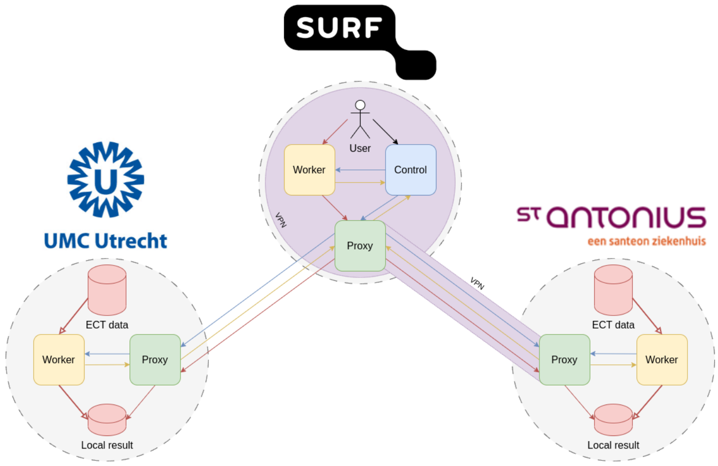


Figure 3 Schematic overview of the PoC setup. Three domains participate in the proof-of-concept, each of which has their own resources hosting their part of the framework: each hospital (UMC Utrecht and St. Antonius) have local instances of the ECT dataset, and host a worker node to perform local computations; SURF, meanwhile, acts as an entry point, and hosts a worker to aggregate the local results into a global one. Domains are contained in each dotted sphere, where each sphere shows the relevant components for that domain. The arrows indicate network traffic between the domains, where the direction of the arrow indicates the direction of the initial message.

per application scenario by utilizing virtualized services hosted by a bridging node. The framework runs a middleware in a virtualized manner and offers network services to secure data sharing and consider policies.

9 The EPI Proof of Concept

In order to experimentally validate the work presented in this paper, a proof-of-concept (PoC) has been deployed in cooperation with various organizations in the EPI consortium. The goal of the PoC is to create an environment in which federated applications may be executed which involve one or more "local" processing steps followed by a centralized "global" step that aggregates the local results. In the PoC, local steps are executed by St. Antonius and UMC Utrecht, each having a part of a horizontally split dataset. To test use-case 3.2, as well as our methods of analysing distributed data at scale (section 5), we use a dataset that has been previously collected in both hospitals with characteristics and treatment outcomes of patients who received electroconvulsive therapy. Results of the local computations on this dataset are sent to SURF, which plays the role of a trusted third-party that can aggregate the results (see figure 3). SURF also hosts the BRANE runtime that acts as an orchestrator.

The PoC is used to assert that the framework discussed in sections 7 and 8 can support the resource pool hosted by the three participating organizations. Specifically, the resources provided are heterogeneous in three of the aforementioned dimensions: they differ in hardware capabilities, software stacks and trust relations. Regarding trust: only the hospitals are allowed to see only their own ECT dataset. SURF is only allowed to see the local results produced on each site, and the global result they produce. These rules are ideal to experiment with automated policy enforcement (section 6). The setup also validates whether the framework can operate within the administrative and security restrictions imposed by the security requirements of the hospitals. Most notably, virtual private networks (VPNs) [42] are used to safeguard the data as it travels over the public network, network access is restriced to the outside world and some domains only offer restricted rights to the framework runtime.

10 Conclusion

Employing digital health twin applications to empower personalized medicine is only feasible utilizing medical data sharing and patient-generated data. On the other hand, data sharing introduces a plethora of challenges ranging from effective data processing modules, to policy compliance, and the underlying heterogenous infrastructure. That is especially true in healthcare, where data privacy and patients' confidentiality is a concern.

In the EPI project, we address these challenges by, focusing on solving three use cases involving consortium members. These serve as exemplary applications of personalized medicine and help us to build applicable solutions. Subsequently, we introduce health data processing solutions (real-time response statistical learning) to provide robust personalized recommendations. Data in the real world is distributed, so we investigate distributed learning and various methods to include privacy preserving techniques. Then, we build policy reasoning tools, where we formalize policies around these use case to control data usage whilst running workflows. Additionally, we implement BRANE; which is a distributed research platform; to build and run said workflows irrespective of the heterogeneous nature of the underlying computing resources. Lastly, we address any network and security discrepancies across the data sharing members by provisioning dynamic network and security services and ultimately enforce any policies in place. This is further showcased by running a proof of concept and utilizing the EPI Framework features by running a data sharing use case across consortium members. The infrastructure for running the PoC was provided by the hospitals to run local functions, and by SURF to manage and orchestrate these services.

Altogether, these result bring us closer to the creation of Health Digitial Twins, a necessary component in enabling personalised interventions. Still the EPI experiences and insights trascend the medical domain. In fact, a major ambition of the EPI project, and in general of the Commit2Data funding scheme to which the project belongs, was to create long term and broadly applicable results. To this aim, we are working to migrate the EPI software framework, which incorporates all project results, to other generic data sharing infrastructures, such as AMdEX.

References

1 Sushant Agarwal, Simon Steyskal, Franjo Antunovic, and Sabrina Kirrane. Legislative compliance assessment: framework, model and gdpr instantiation. In *Privacy Technologies and Policy: 6th Annual Privacy Forum, APF 2018, Barcelona, Spain, June 13-14, 2018, Revised Selected Papers 6*, pages 131–149. Springer, 2018. `doi:10.1007/978-3-030-02547-2_8`.

2 Corinne G. Allaart, Björn Keyser, Henri Bal, and Aart van Halteren. Vertical split learning - an exploration of predictive performance in medical and other use cases. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. `doi:10.1109/IJCNN55064.2022.9891964`.

3 Joshua Baugh, Ute Bartels, James Leach, Blaise Jones, Brooklyn Chaney, Katherine E Warren, Jenavieve Kirkendall, Renee Doughman, Cynthia Hawkins, Lili Miles, et al. The international diffuse intrinsic pontine glioma registry: an infrastructure to accelerate collaborative research for an orphan disease. *Journal of neuro-oncology*, 132(2):323–331, 2017.

4 Giovanni Briganti, Marco Scutari, and Richard J McNally. A tutorial on bayesian networks for psychopathology researchers. *Psychological methods*, 2022.

5 Koen Bruynseels, Filippo Santoni de Sio, and Jeroen van den Hoven. Digital twins in health care: Ethical implications of an emerging engineering paradigm. *Frontiers in Genetics*, 9, 2018. `doi:10.3389/fgene.2018.00031`.

6 Iker Ceballos, Vivek Sharma, Eduardo Mugica, Abhishek Singh, Alberto Roman, Praneeth Vepakomma, and Ramesh Raskar. Splitnn-driven vertical partitioning. *arXiv preprint arXiv:2008.04137*, 2020. arXiv:2008.04137.

7 Council of the EU. General Data Protection Regulation. *Official Journal of the European Union*, 59, 2016.

8 Ewa Deelman, Karan Vahi, Mats Rynge, Rajiv Mayani, Rafael Ferreira da Silva, George Papadimitriou, and Miron Livny. The evolution of the pegasus workflow management software. *Computing in Science & Engineering*, 21(4):22–36, 2019. doi:10.1109/MCSE.2019.2919690.

9 Vanessa Díaz-Zuccarini and Silvia Schievano. Biomedical imaging and computational modeling in cardiovascular disease: Patient-specific applications using numerical models. *Biomedical Imaging and Computational Modeling in Biomechanics*, pages 173–192, 2013.

10 Celtia Domínguez-Fernández, June Egiguren-Ortiz, Jone Razquin, Margarita Gómez-Galán, Laura De las Heras-García, Elena Paredes-Rodríguez, Egoitz Astigarraga, Cristina Miguélez, and Gabriel Barreda-Gómez. Review of technological challenges in personalised medicine and early diagnosis of neurodegenerative disorders. *International Journal of Molecular Sciences*, 24(4), 2023. doi:10.3390/ijms24043321.

11 Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

12 Glyn Elwyn, Dominick Frosch, Richard Thomson, Natalie Joseph-Williams, Amy Lloyd, Paul Kinnersley, Emma Cording, Dave Tomson, Carole Dodd, Stephen Rollnick, et al. Shared decision making: a model for clinical practice. *Journal of general internal medicine*, 27:1361–1367, 2012.

13 Christopher A. Esterhuyse, Tim Müller, L. Thomas Van Binsbergen, and Adam S. Z. Belloum. Exploring the enforcement of private, dynamic policies on medical workflow execution. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 481–486, 2022. doi:10.1109/eScience55777.2022.00086.

14 Peter Grünwald. Beyond neyman-pearson. *arXiv preprint arXiv:2205.00901*, 2022.

15 Qijian He, Wei Yang, Bingren Chen, Yangyang Geng, and Liusheng Huang. Transnet: Training privacy-preserving neural network over transformed layer. *Proceedings of the VLDB Endowment*, 13(12):1849–1862, 2020. URL: http://www.vldb.org/pvldb/vol13/p1849-he.pdf.

16 Lindsey M Hoffman, Sophie EM Veldhuijzen Van Zanten, Niclas Colditz, Joshua Baugh, Brooklyn Chaney, Marion Hoffmann, Adam Lane, Christine Fuller, Lili Miles, Cynthia Hawkins, et al. Clinical, radiologic, pathologic, and molecular characteristics of long-term survivors of diffuse intrinsic pontine glioma (dipg): a collaborative report from the international and european society for pediatric oncology dipg registries. *Journal of clinical oncology*, 36(19):1963, 2018.

17 Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

18 Renato Iannella and Serena Villata. Odrl information model 2.2. *W3C Recommendation*, 15, 2018.

19 MHA Jansen, DG Van Vuurden, WP Vandertop, and GJL Kaspers. Diffuse intrinsic pontine gliomas: a systematic update on clinical trials and biology. *Cancer treatment reviews*, 38(1):27–35, 2012.

20 Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, and Keqiu Li. Big data processing in cloud computing environments. In *2012 12th international symposium on pervasive systems, algorithms and networks*, pages 17–23. IEEE, 2012.

21 Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021. doi:10.1561/2200000083.

22 Jamila Alsayed Kassem, Adam Belloum, Tim Müller, and Paola Grosso. Utilisation profiles of bridging function chain for healthcare use cases. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 475–480, 2022. `doi:10.1109/eScience55777.2022.00085`.

23 Jamila Alsayed Kassem, Cees De Laat, Arie Taal, and Paola Grosso. The epi framework: A dynamic data sharing framework for healthcare use cases. *IEEE Access*, 8:179909–179920, 2020. `doi:10.1109/ACCESS.2020.3028051`.

24 Jamila Alsayed Kassem, Onno Valkering, Adam Belloum, and Paola Grosso. Epi framework: Approach for traffic redirection through containerised network functions. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 80–89, 2021. `doi:10.1109/eScience51609.2021.00018`.

25 Milen G Kebede, Giovanni Sileno, and Tom Van Engers. A critical reflection on odrl. In *AI Approaches to the Complexity of Legal Systems XI-XII: AICOL International Workshops 2018 and 2020: AICOL-XI@ JURIX 2018, AICOL-XII@ JURIX 2020, XAILA@ JURIX 2020, Revised Selected Papers XII*, pages 48–61. Springer, 2021. `doi:10.1007/978-3-030-89811-3_4`.

26 Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

27 Shilpa Krishnan, Catherine C Hay, Monique R Pappadis, Anne Deutsch, and Timothy A Reistetter. Stroke survivors' perspectives on post–acute rehabilitation options, goals, satisfaction, and transition to home. *Journal of Neurologic Physical Therapy*, 43(3):160–167, 2019.

28 Peter Langhorne, Julie Bernhardt, and Gert Kwakkel. Stroke rehabilitation. *The Lancet*, 377(9778):1693–1702, 2011. `doi:10.1016/S0140-6736(11)60325-5`.

29 Guido Noordende, Silvia Olabarriaga, Matthijs Koot, and Laat M. Trusted data management for grid-based medical applications, January 2011.

30 Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018. `arXiv:1802.08908`.

31 Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health information science and systems*, 2(1):3, 2014. `doi:10.1186/2047-2501-2-3`.

32 Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*, 2022. `doi:10.48550/arXiv.2210.01948`.

33 OASIS Standard. extensible access control markup language (xacml) version 3.0. *A:(22 January 2013). URl: http://docs. oasis-open. org/xacml/3.0/xacml-3.0-core-spec-os-en. html*, 2013.

34 Judith ter Schure and Peter Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, 2019.

35 Rosanne J. Turner, Femke Coenen, Femke Roelofs, Karin Hagoort, Aki Härmä, Peter D. Grünwald, Fleur P. Velders, and Floortje E. Scheepers. Information extraction from free text for aiding transdiagnostic psychiatry: constructing nlp pipelines tailored to clinicians' needs. *BMC psychiatry*, 22(1):407, 2022.

36 Rosanne J. Turner and Peter D. Grünwald. Safe sequential testing and effect estimation in stratified count data. *arXiv preprint arXiv:2302.11401*, 2023.

37 Rosanne J. Turner and Peter D. Grünwald. Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, page 109835, 2023. `doi:10.1016/j.spl.2023.109835`.

38 Onno Valkering, Reginald Cushing, and A. Belloum. Brane: A framework for programmable orchestration of multi-site applications. In *17th IEEE International Conference on eScience, eScience 2021, Innsbruck, Austria, September 20-23, 2021*, pages 277–282, September 2021. `doi:10.1109/eScience51609.2021.00056`.

39 L Thomas van Binsbergen, Milen G Kebede, Joshua Baugh, Tom Van Engers, and Dannis G van Vuurden. Dynamic generation of access control policies from social policies. *Procedia Computer Science*, 198:140–147, 2022. doi:10.1016/J.PROCS.2021.12.221.

40 L. Thomas van Binsbergen, Lu-Chi Liu, Robert van Doesburg, and Tom van Engers. Eflint: A domain-specific language for executable norm specifications. In *Proceedings of the 19th ACM SIGPLAN International Conference on Generative Programming: Concepts and Experiences*, GPCE 2020, pages 124–136, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3425898.3426958.

41 Lourens E. Veen, Sara Shakeri, and Paola Grosso. Mahiru: a federated, policy-driven data processing and exchange system, 2022. arXiv:2210.17155, doi:10.48550/arXiv.2210.17155.

42 R. Venkateswaran. Virtual private networks. *IEEE Potentials*, 20(1):11–15, 2001. doi:10.1109/45.913204.

43 Eric-Jan Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, 14(5):779–804, 2007.

44 Qingchen Zhang, Laurence T. Yang, Zhikui Chen, and Peng Li. A survey on deep learning for big data. *Information Fusion*, 42:146–157, 2018. doi:10.1016/j.inffus.2017.10.006.

45 Tianwei Zhang, Zecheng He, and Ruby B Lee. Privacy-preserving machine learning through data obfuscation. *arXiv preprint arXiv:1807.01860*, 2018. arXiv:1807.01860.

Helping Cancer Patients to Choose the Best Treatment: Towards Automated Data-Driven and Personalized Information Presentation of Cancer Treatment Options

Emiel Krahmer[1]
Department of Communication and Cognition, Tilburg University, The Netherlands

Felix Clouth
Department of Methodology and Statistics, Tilburg University, The Netherlands
Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands

Saar Hommes
Department of Communication and Cognition, Tilburg University, The Netherlands
Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands

Ruben Vromans
Department of Communication and Cognition, Tilburg University, The Netherlands
Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands

Steffen Pauws
Department of Communication and Cognition, Tilburg University, The Netherlands
Innovation Excellence, Innovation & Strategy, Philips, Eindhoven, The Netherlands

Jeroen Vermunt
Department of Methodology and Statistics, Tilburg University, The Netherlands

Lonneke van de Poll-Franse
Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands
Department of Medical and Clinical Psychology, Tilburg University, The Netherlands
Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

Xander Verbeek
Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, The Netherlands

Abstract

When a person is diagnosed with cancer, difficult decisions about treatments need to be made. In this chapter, we describe an interdisciplinary research project which aims to automatically generate personalized descriptions of treatment options for patients. We relied on two large databases provided by the Netherlands Comprehensive Cancer Organisation (IKNL): The Netherlands Cancer Registry and the PROFILES dataset. Combining these datasets allowed us to extract personalized information about treatment options for different types of cancer. In a next step we provided personalized context to these numbers, both in verbal statements and in narratives, with the aim to facilitate shared decision making about treatments. We discuss strengths and limitations of our approach, illustrate how it generalizes to other health domains, and reflect on the overall research project.

2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI); Computing methodologies → Artificial intelligence; Applied computing → Life and medical sciences

[1] Corresponding author

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 3; pp. 3:1–3:20
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Keywords and phrases Oncology, Data-driven shared decision making, Latent class analysis, Risk communication, Narratives, Personalization

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.3

Funding This research was funded by the commit2data program from the Netherlands Organisation for Scientific Research (NWO) [Grant No. 628.001.030], by the Data Science Center Tilburg (DSC/t), by the "We Care" program from Tilburg University and Elisabeth-TweeSteden hospital [Grant No. 42001155], and by the Netherlands Organization for Health Research and Development (ZonMW), "Kennisvragen Uitkomstgerichte Zorg" [Project No. 80-83900-98-3003], which are all gratefully acknowledged.

1 Introduction

In his book *Mortality*, the British author Christopher Hitchens memorably describes how following his cancer diagnosis he found himself transported "from the country of the well across the stark frontier that marks off the land of malady" [34]. Every year many other people experience the same transition, and have to adjust to the medical, emotional, and cognitive challenges of their new habitat. One of the first of these challenges involves making informed decisions about possible treatments. Depending on the form and stage of the cancer, there may be a range of treatment options, each resulting in different chances for survival, risks of side effects and general impact on quality of life during and after treatment.

For example, in the case of prostate cancer, treatment options include (but are not limited to) radiotherapy, prostatectomy (removal of part or all of the prostate gland) and active surveillance (monitoring the cancer closely, hoping to avoid invasive treatment) [28]. Interestingly, these treatments have highly comparable survival rates. Hamdy et al. [28] report about 1 death per 1000 person-years for each of these treatments, and also find no significant differences between the treatment groups in the number of deaths from any cause. However, the treatments differ markedly in terms of potential impact on quality of life and side effects, which may range from incontinence or erectile dysfunction, to fear or anxiety [17].

In general, and in the ideal case, doctor and patient (often in concordance with loved ones) jointly decide upon the treatment plan.[2] However, such "shared decision-making" [19] is not always easy, and crucially relies on patients being adequately informed about favorable outcomes (such as long-term survival) and risks of adverse effects (such as side effecs) of potential treatments.

In short, patients can only make an informed choice about treatment and thereby participate in shared decision-making, when they are properly informed. This, however, is not easy. Oncologists generally face difficulties in how to inform their patients, because decisions typically hinge on weighing chances, which are not always easy to access, understand, and explain to patients [26]. Often the benefits and risks of different options for an individual patient are not precisely known in advance, and the statistics are derived from general prediction models, which are based on existing data of earlier cases (e.g., [20]). When doctors want to communicate this kind of information to patients, they can make use of information leaflets or websites, and sometimes also of decision aids. The latter are tools that can help patients to play a role in decision making, together with their doctor, by making the decision points explicit, and linking them to options and possible outcomes [7].

[2] A note about terminology: we use the term 'cancer survivor' for anyone who has had a cancer diagnosis and is still alive, while we reserve the term "patient" for anyone who is currently preparing for or receiving treatment.

However, there are three major challenges with this general scenario:

1. There is a focus on survival, with less attention to other factors. Obviously, long term (preferably recurrence-free) survival is a crucial factor for cancer treatment decision making. But, arguably, it is not the only one that should be taken into account. Side-effects of treatment, quality of life after treatment, and general well-being are example factors that patients increasingly consider to be very important as well. These factors will only become more relevant when cancer is increasingly likely to be viewed as a chronic disease [55].

2. Information is often collected and presented on a population level, rather than personalized (e.g., [38, 60, 67]). Still, it is intuitively clear that personalizing such information increases the relevance. Consider the case of prostate cancer again: it makes a big difference whether one receives a first diagnosis at 50 or at 80 years old, both in terms of the progression of the disease and of the potential impact of possible treatment side effects. In this situation, personalized information is clearly much more relevant.

3. Good and clear communication is a *conditio sine qua non* for information presentation about treatment options. However, it is difficult to get this right, especially when information is personalized, which creates two new challenges: (a) It is not obvious how personalized risks and opportunities should be explained to patients, nor to what extent they will be able to understand them. (b) Moreover, when information is personalized, this also means that communication should be different for each individual person, and this is clearly not something that can be obtained with a static patient information leaflet. Hence, this calls for a new approach.

In our research project, we tackled these three related challenges in an interdisciplinary approach, in which statisticians and epidemiologists collaborated with AI-researchers, computational linguists and communication scientists. Additionally, we intensively collaborated with the Netherlands Comprehensive Cancer Organisation (IKNL). IKNL has been actively involved from the very initial stages of the project, helping write the proposal, co-defining the research questions and generally helping to shape and conduct the research we undertook. Crucially, IKNL also provided two data-sets, which we relied on as the foundation of the project: The Netherlands Cancer Registry [44] and the PROFILES dataset [56]. Since 1989, the NCR tracks all new cancer diagnoses in The Netherlands, with data about diagnosis and treatment collected directly from hospitals. It currently contains data of over 2,500,000 persons. PROFILES is a registry for the study of physical and psychosocial impact of cancer and its treatment from a population-based cohort of cancer survivors. Since 2004, over 25,000 people have participated, reporting the consequences of cancer and its treatment on perceived quality of life, symptoms, and societal participation. Combining these datasets allowed us to extract personalized information about treatment options for different types of cancer (as we describe in Section 2 below). This enables us to determine, say, that a patient with cancer-type W and demographic factors X has a probability Y of experiencing treatment side-effect Z. But how do we communicate this to the patient? Do we use words or numbers? Do patients actually want and understand this kind of personalized information? This is addressed in Section 3. Besides describing statistics, it is important to also give personalized context to these numbers, since we can anticipate that merely communicating numbers will not be perceived as helpful and persuasive by everyone. Hence we also explore, in Section 4, whether providing personalized information in the form of tailored narratives is feasible and appreciated. We end this paper, in Section 5, with a general discussion of lessons learned and outstanding issues.

2 Computing personalized information

Computing personalized predictions on quality of life (QoL) for cancer survivors poses a multitude of interesting challenges. Many prediction models have been developed to support decision-making in oncology [30, 57] and there is a large amount of resources and methodological guidance to support the development of new prediction models. However, computing personalized predictions on QoL poses a rather unique problem compared to existing prediction models. Most prediction models have a clearly defined, directly observable outcome, e.g., additional time of survival [68], recurrence-free survival [8], or recurrence of the tumor [9]. In contrast, QoL is neither clearly defined nor directly observable [40]. If we were to ask ten people what their QoL is we would get ten different answers. Not necessarily because their QoL is very different but because they were thinking about different aspects of QoL when providing their answers. While the first person might have been thinking about the physical struggles to carry home a bag of groceries, the second person was happy to feel less pain than last month, and the third person was thankful for still being able to play with his or her grandchildren.

Ideally, we want to account for all aspects of QoL when constructing a prediction model. For this to be possible, we need high-quality data that contains information about all of the aspects of QoL. This requires tedious data collection using questionnaires, often at a very difficult time in the survivors' lives. Thankfully, the PROFILES study [56] offers data of such patient-reported outcome measures (PROMs) for cancer survivors with a wide range of diagnoses and treatments. Furthermore, the data from the PROFILES study can be directly linked to the Netherlands Cancer Registry (NCR [44]). The rich information from this register allows access to all relevant medical information including treatments of the survivors for whom we have PROMs data. The PROFILES study uses validated questionnaires such as the EORTC QLQ-C30, a comprehensive tool for assessing QoL that accounts for its multidimensionality, and its data has been widely used for research on cancer survivorship (for an overview, see [49]).

Subsequently, the model chosen to predict QoL should reflect the multidimensionality of this construct [11, 40]. Rather than developing a prediction model for each dimension separately, our goal was to predict one single outcome that entails all the information of the multiple dimensions of QoL. Keeping in mind the aim of using the predictions in decision aids, instead of providing explanations for predicted scores on each (for some patients potentially uninteresting) dimension of QoL, we aimed for a single outcome that could more easily be communicated to patients (described in Section 3), or even be translated into narratives (Section 4 further elaborates on this use).

To achieve this, we chose to use latent class analysis (LCA), a model-based clustering algorithm. Based on a set of indicators, in this case the dimension scores of QoL, LCA identifies distinct latent classes of cancer survivors with similar patterns of QoL. These QoL classes can then be interpreted based on the (average) values on the QoL items for the survivors in the respective classes. For example, in Clouth et al. [11], one of the five identified classes was characterized by very good scores on all QoL dimensions (class 1; Figure 1), while another class was characterized by very good scores on the social, cognitive, and emotional functioning scales but showed limitations on the physical and role functioning, fatigue, and pain dimensions (class 4), and a third class had moderate to good scores on the (short-term) side effect symptoms but low scores on the more long-term functioning dimensions (class 5).[3]

[3] Note that the identification of the best model in terms of number of classes is difficult. Several fit

LCA then estimates the probabilities of membership in these classes for each individual. The more comprehensive interpretations of the classes are more graspable and (potentially) more interesting for patients. Furthermore, these interpretations can directly be used as input for narratives in decision aids.

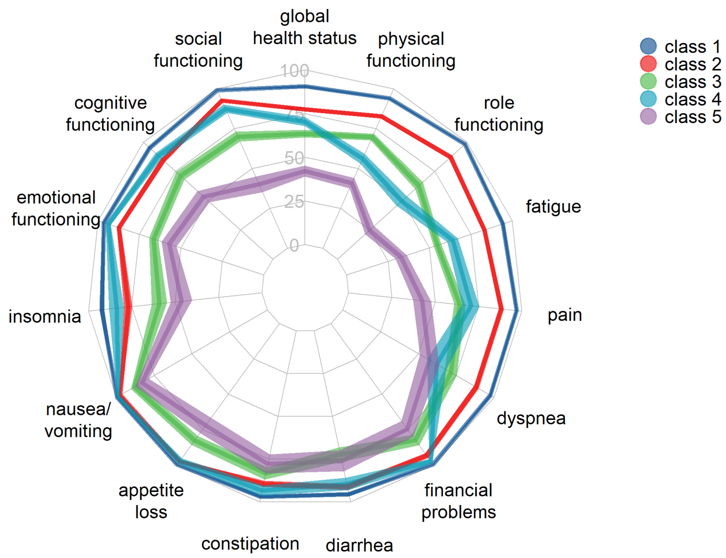


Figure 1 Means and confidence intervals of all five health-related quality of life classes for the 15 European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire C30 dimensions. On the upper half are the functioning scales and on the bottom half are the symptom scales of (short-term) side effects related to treatment. The symptom and single-item scales were reversed.

After identifying the outcome of interest, decision-making requires relevant and personalized information [19]. When choosing a specific treatment option, organizing follow-up care, and thinking about lifestyle arrangements, predictions based on average QoL can give some indication. However, the real interest lies in personalized predictions, that is, what is the expected QoL for "someone like me".[4] Statistically speaking, "someone like me" refers to individuals of similar age, gender, socioeconomic status, general (physical) health, and similar diagnosis in terms of location, size, and aggressiveness of the tumor. These predictors were included in a regression model to estimate the expected probabilities for membership in the QoL classes conditional on the values of the predictors [11]. With the estimated model

indices are available to support decision making but interpretability of the obtained solution is often regarded as equally important.

[4] Additional to personalized predictions, benchmarks might also be important to patients as they often want to know how they are doing compared to others. However, such benchmarks are most often well known and readily available to the treating doctors. We therefore did not focus on this aspect in our project.

parameters, we can then in turn calculate the expected score of class membership probability for every new patient for whom we have the relevant information on the predictors. For instance, imagine a new patient X, who is a 47 years old male with stage 2 colon cancer, no comorbidities, and who received chemotherapy additional to surgery. For this patient, we would then compute a large probability (68%) of being part of class 1 (excellent QoL), and a smaller, but still substantial probability (25%) to be in class 2 (good QoL with prevalence of insomnia). Noticeably, this patient would have a very small probability (2%) of being in class 5 (poor QoL with severe limitations). We deliberately chose for LCA in combination with regression techniques for estimating the personalized predictions to account for the uncertainty around our predictions. While there are many algorithms for clustering and prediction available, parametric approaches such as LCA and regression have the advantage that information about uncertainty, in particular the standard errors of the estimates, are obtained implicitly.

Predictions like the ones described above can be greatly beneficial for patients to support decision-making after diagnosis. However, one aspect of this decision-making, the decision about treatment, requires additional considerations. The question of which treatment option is preferable is a classical "what if" question, that is, a question concerning the causal effect of the treatment [32]. While prediction is aimed at forecasting observations under known conditions, causal inference is aimed at making predictions under unknown conditions. That is, the prediction of the effect of a treatment option A for a specific patient is only valid if that specific patient would actually receive treatment A. The problem is that if there is a good reason for this patient to receive treatment B, only the prediction of the effect of treatment option B (but not of treatment option A) is valid. Since we will only ever observe one of these scenarios, we cannot predict how much better treatment option A is for this specific patient. To arrive at the answer for this, we need to ensure that the patients under treatment A are equal (or similar enough) to the patients under treatment B (exchangeability assumption [32]). While this is usually achieved using randomized controlled trials, it is a major challenge when using observational data (such as registry data). Several causal inference techniques have been developed to overcome these challenges [32]. Unfortunately, these techniques are not directly applicable to be used in combination with clustering algorithms such as LCA. In Clouth et al. [12] and Clouth et al. [13], we developed new methods that combine these established causal inference techniques with LCA. In particular, we show how inverse propensity weighting (IPW) can be included in LCA to estimate the causal effect of surgery compared to an active surveillance strategy for low-stage prostate cancer patients. After estimating a measurement model for the latent QoL classes, the IPW weights can be combined with weights for the measurement errors from this first model. These new weights can then be included in a regression model to estimate the causal effect. In Clouth et al. [12], we show that there is no causal effect of surgery compared to an active surveillance strategy on the probability of class memberships in either good or bad QoL classes.

The impact of the methodological work in this project is beyond the domain of oncology and survivorship research. We validated the methodology of using a combination of LCA and regression techniques as well as causal inference techniques on several other data sets, both cross-sectional and longitudinal. The examples range from Australian data on breast cancer survivors [15], to data on trauma patients admitted to the ICU [29], cohort data from the US on drug abuse [13], Dutch panel data on mental health [10], and data on self-awareness and autonomy [42].

3 Explaining personalized information

A critical question that needs to be addressed is whether patients really want to receive personalized statistics about treatment outcomes obtained via prediction models. Should we disclose such statistics, and if so, how and to whom? By conducting both qualitative [37, 36, 66] and quantitative [61] studies, we found that the vast majority of patients and survivors, especially those with higher subjective numeracy skills and an information-seeking coping style, have a desire to receive personalized statistics about risks of treatment side effects, survival, and quality of life. Patients' considerations for wanting personalized cancer statistics related to feelings of being in control or making better-informed decisions, while considerations for not wanting statistics were about the unpredictability of future events for individual patients or negative experiences with statistics in the past. Personalized statistics were also considered more useful and relevant compared to generic, non-personalized statistics. Importantly, and in line with related research (e.g., [23, 58, 70]), whether personalized risks should be communicated to patients strongly depends on individual differences. For instance, those who are not actively seeking for detailed information and who are subjectively less numerate may have less desire for received personalized statistics information about treatment outcomes [48].

Once we established that patients have a need for receiving personalized statistical information, we wanted to explore what this would add to clinical practice. In two qualitative studies, we interviewed patients (N=35) and clinicians (N=6) about the role of personalized statistics in clinical practice. What is missing now and how could clinical practice benefit from personalized statistical information? We found that patients have a general "gist" knowledge about personalized predictions, they understand for example that their outcomes would improve over time and that some outcomes might take a while before they improve [6, 37, 36]. It seemed more difficult however, to understand the prediction model, the input data, and the statistical procedures that were responsible for the personalized predictions. Figure 2 shows an example of how the (input to the) prediction model was visualized. Even without precise knowledge on the working of the prediction model, receiving predictions reassured patients and most liked knowing what life would look like after their disease – especially if their prospects were not too positive. Some patients were not that worried about life after their diagnosis so they experienced a lesser need for detailed information. The most important implication for clinical practice would be that patients indicated they would initiate conversations with their doctor more about how they were doing and they would think about what to say during a consultation beforehand.

However, merely providing patients with personalized data is not enough. These numbers need to be effectively and adequately communicated to help patients make sense of their personalized health information. Therefore, we conducted several pre-registered experiments among a sample of cancer patients and survivors (N=141) and a Dutch representative sample of healthy participants (N=1807) with the aim to examine how varying message formats and contextual strategies for communicating personalized risks can influence people's cognitive, emotional, and behavioral responses. When presenting personalized risks to patients, they can best be communicated numerically using natural frequencies (e.g., "40 out of 100 men like you will experience this side-effect") combined with verbal descriptors (e.g., "this risk is common") and/or visual displays such as icon arrays [62, 66]. As illustrated in Figure 3, verbal-only formats (without any numerical data) should be avoided, as they may lead to variable and inaccurate risk perceptions (for similar reasoning, see [5]).

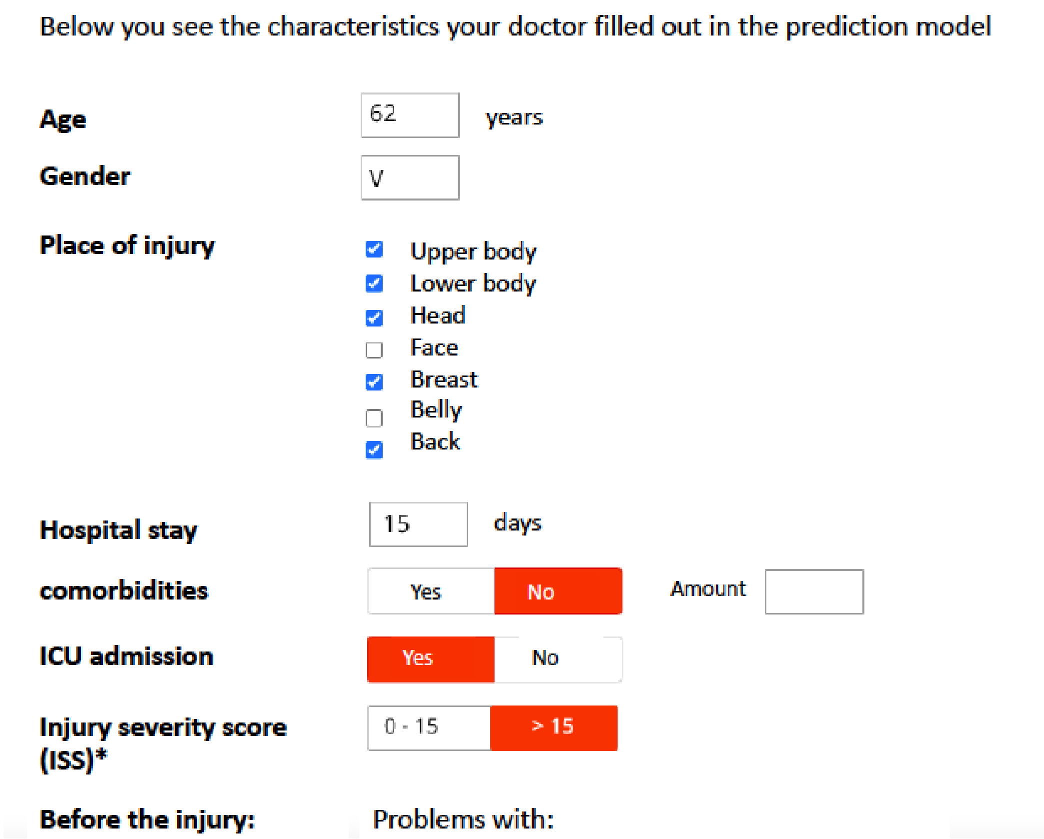


Figure 2 Example of the way the input data for the prediction model could be visualized. This example was used for trauma patients (people who had been in an accident) and was based on the prediction model developed in [29].

Moreover, even single risk statistics that are personalized and presented in isolation may be difficult to evaluate by patients and are sometimes even ignored by them [21, 72]. Results from both needs assessments and an experimental study show that people generally want comparative data about the average person's risk or survival rate when being provided with their personalized risk. By doing so, patients are better able to interpret even unfamiliar personalized risks because both types of risks serve as a reference for each other, which helps patients determine the "goodness" and the "badness" of their personalized health data [71]. Our studies show that providing such comparative data (Figure 4) increases information evaluability and can help patients derive meaning from their personalized risk data, without negatively influencing their risk perceptions, emotional responses or behavioral treatment intentions [63]. In addition, providing simple explanations about how personalized risks are determined (e.g., by discussing risk factors) may help patients better understand and recall those risks, and further suggest that they can play an important role in shared decision-making about treatment [64].

4 Narrating personalized information

One of the challenges in communicating statistics to patients is that people might not always correctly interpret numbers [26]. This lack of numeracy skills might in turn influence their perceived usefulness or negatively impact health-related decisions [31]. One way of

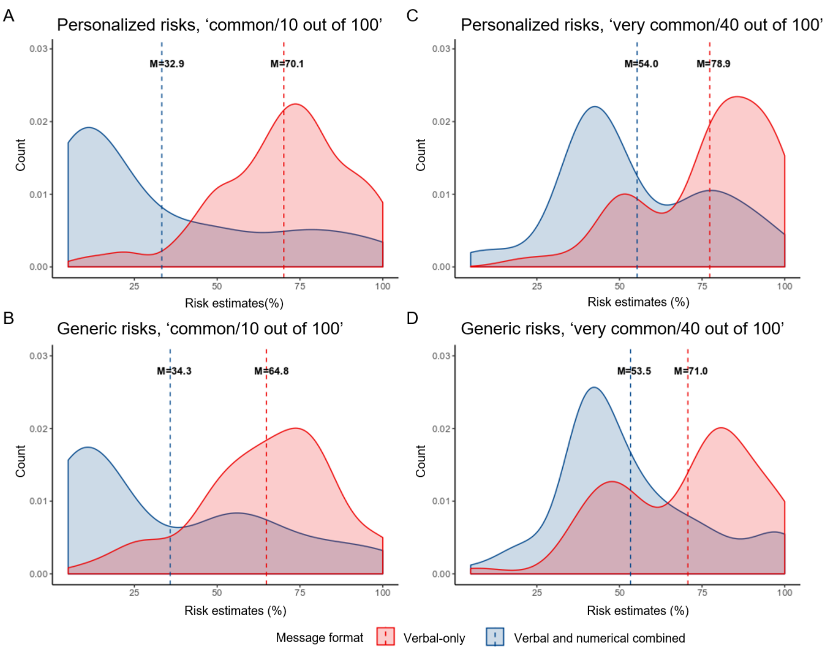


Figure 3 Variation in risk estimates given by patients when being presented with personalized and generic risks presented via verbal descriptors (e.g., "this side effect is common") with or without numbers ("occurs in 10 out of 100 patients like you").

overcoming this challenge might be by contextualizing numbers in stories, often called narratives. Narratives can be defined as "a representation of connected events and characters that has an identifiable structure, is bounded in space and time, and contains implicit or explicit messages about the topic being addressed" ([41], p.222). In health communication, narratives have been extensively studied in persuasive contexts where readers need to be persuaded to make healthier choices. They have proven effective tools in trying to persuade people to eat healthier, attend screening or quit smoking for example [54]. Narratives might also lead to better recall, fewer counterarguments and experiencing more positive emotions [52].

In informative settings, the setting of our project, the role of narratives is less straightforward. Whether or not narratives are helpful in informing readers might be measured by a range of variables including for example better recall, knowledge increase or risk estimation. In these informative settings, where narratives need not persuade readers, there is worry for the narrative bias effect [14, 39, 53]. Narrative bias entails that a narrative is so powerful that it outweighs statistical or factual information, even when the narrative counters the statistical information. It seems that this is especially the case when narratives are about adverse effects of health decisions, such as a narrative portraying severe side effects after a vaccine [27]. To combat narrative bias and to still make use of its positive effects, it is arguably more fruitful to develop narratives that are in line with the statistical information rather than narratives that counter the facts. That is why in our final studies, we used balanced narratives that are in line with the statistical information.

One way of making narratives more appealing and relevant for readers is by personalizing them. Personalizing (or tailoring) health information has shown promising results. Personalization means that the narrative or the character in the narrative is matched to the reader of

Figure 4 Providing comparative data about the average person's risk may help patients better evaluate their personalized risk data.

that narrative. To give an example, Baezconde-Garbanati et al. [2] showed that a narrative personalized to the cultural background of readers was more effective in persuading readers to undergo cervical cancer screening than a narrative that was not personalized to the reader. In addition, a systematic review by Noar et al. [45] found that in general, personalized health information leads to better health outcomes than generic information. Additionally, studies found that the first-person perspective leads to better effects than other perspectives [51].

A downside to personalized health information however may be that it is a time-consuming and labor-intensive process if all information should be personalized based on individual characteristics of the reader. That is why we call for automatic personalization rather than manually changing health information [4, 47]. By applying basic data-to-text techniques, we are now able to construct a narrative that is personalized based on individual patient characteristics. Figure 5 shows examples of narratives we used in our study and how they were personalized depending on the reader.

David (76) was diagnosed with melanoma in 2020.

"I discovered an odd spot on my leg. At first, I thought it was nothing, but just to be sure I went to my general practitioner. They made a biopsy in the hospital and it turned out I had skin cancer. Of course, I know that at my age I can become ill, but still I was quite shocked."

As a part of his routine care, David regularly filled out questionnaires. He could see his results so he could form an idea of how his symptoms might change. For example, about physical functioning:

"I didn't really know what I could expect after the treatment, would I for example still be able to enjoy retirement and go on a vacation? At first, I didn't know if I wanted to know my own scores from the questionnaire on physical functioning, but it helped me to answer these questions. When I saw that my results were better than the critical value, it did reassure me. Of course, that doesn't mean that I didn't feel anything at all from the treatment, but it could have been worse. With small adjustments I could go for a walk and actively hang out with the grandkids again."

[...]

Outcome	Your score	Critical value
Physical functioning	87	83
Role functiong	100	58

Hanneke (36) was diagnosed with bladder cancer in 2020.

"I discovered blood in my urine. At first, I thought it was nothing, but just to be sure I went to my general practitioner. After a blood test they made a biopsy in the hospital, and it turned out I had bladder cancer. Because I am still young, I really did not see the diagnosis coming."

As part of her routine care, Hanneke regularly filled out questionnaires. She could see her results so she could get an idea of how her symptoms might change. For example, about physical functioning:

"I didn't really know what I could expect after the treatment, would I for example still be able to carry out my busy job and exercise? At first, I didn't know if I wanted to know my own scores from the questionnaire on physical functioning, but it helped me to answer these questions. When I saw that the results were worse than the critical value, that confirmed what I already suspected. I talked to my doctor about it because I wanted to know what I could do to be able to exercise again and go on vacation."

[...]

Outcome	Your score	Critical value
Physcial functioning	50	83
Role functioning	33	58

Figure 5 The first paragraphs of two personalized narratives from fictional patients 'David' and 'Hanneke'. Narratives are personalized based on the reader's characteristics. These included: gender (male/female), age (old, middle, young), diagnosis (skin, bladder, breast cancer), physical functioning score (better or worse than critical value) and role functioning (better or worse than critical value). Additionally, patients could view their personal scores in a table at the end of the narrative. The photos were automatically generated from a website (`https://generated.photos/`). Narratives were constructed based on previously conducted interviews about patient experiences with cancer and were developed together with an expert team including clinicians, communication and data experts.

5 General Discussion and Conclusion

In the foregoing, we have described our novel interdisciplinary framework with the aim of better informing cancer patients about treatment options. By combining different data sources, such as the Netherlands Cancer Registry and the PROFILES dataset, we were able to compute personalized profiles on treatment side-effects and general quality of life after treatment, for patients newly diagnosed with cancer, using latent class analysis. On the basis of these profiles, we could generate personalized information about a range of associated outcomes of different treatment options to patients. These were communicated via various formats, ranging from words and numbers to personalized narratives, in which case both the information and its presentation can be tailored towards the reader. Evaluation studies revealed that which formats are preferred and work best for information presentation is dependent on personal and contextual features. Our approach towards personalized information on treatment options has been evaluated in the oncology domain, but can be generally applied to other healthcare contexts (as we have shown for trauma care). In this final section, we discuss a number of challenges and outstanding issues.

5.1 Challenges when using registry data

Personalized predictions are based on data and the quality of these predictions is inherently linked to the quality of the data. The NCR collects data about diagnosis and treatment directly from hospitals for all new cancer diagnoses in The Netherlands since 1989. Furthermore, the PROFILES study complements the information in the NCR with PROMs data on quality of life for a subset of the patients and survivors in the registry. The direct linkage between these two data sources allows for extracting all the information from the NCR when developing prediction models for QoL. On the other hand, using retrospective registry data is also restrictive.

In particular, we encountered three challenges.

1. Registry data is observational, that is, this data does not contain experimental interventions of any kind. Information is observed for treatments that are administered based on the survivors' characteristics (rather than randomly) and outcomes are observed as a consequence of all of these influencing factors rather than as a consequence of treatment alone. This indication bias, otherwise known as confounding by indication or self-selection into treatment bias, prohibits direct conclusions about the effectiveness of certain treatment options. As discussed in Section 2, several causal inference methods have been developed to estimate causal effects from observational data. Using these methods allows us to give answers to the "what if" questions about which treatment option is preferable.
2. Information from registry data might be outdated to address decisions today. Registry data is retrospective; information on patients is collected over many years and patients might have been treated years ago with methods that might be considered outdated by today's standards. For example, in many cases, lymph node dissections are no longer performed or chemotherapy is replaced by immune therapy. Consequentially, outcome information of recent treatments is often limited in registry data. Evidence-based predictions about such recent treatments are then not possible using registry data.
3. Registry data contains survivorship bias (or immortal time bias). Patients for whom a treatment is registered survived at least until they received such treatment. The information about treatments and their effectiveness is thus based on the subset of survivors among all patients. This problem intensifies when considering outcomes such as QoL as these outcomes are of interest and registered some time (often years) after treatment. The information about QoL is thus based on an even smaller subset of patients that survived treatment and some time after. Put differently, we will never observe QoL for patients who did not survive. However, as this set of patients is still present at the time of decision-making about the treatment, predictions based on the subset of survivors will be biased (survivorship bias). A solution for this problem might be the use of baseline assessments of QoL. Baseline QoL has been shown to often be predictive of QoL after treatment and such information is increasingly often being collected. As QoL is always conditional on survival, this information is still very valuable, for example, to describe the current state of cancer survivorship in the Netherlands. However, as input for decision-making about the treatment, the potential of this data is limited.

As a consequence, we draw two conclusions: (1) Predictions for QoL based on data from long-term survivors are often inadequate for decision-making about treatment. And (2): we need to base predictions for long-term QoL on the observed treatment and these predictions are always conditional on survival. Facing these limitations, we need to stress that these predictions (predictions conditional on survival and based on the observed treatment) can still be tremendously useful for cancer survivors. A cancer diagnosis changes a person's life fundamentally and even after important decisions about the treatment have been made, there will still be a large amount of uncertainty about the future. Tools like the ones described above can play an important role in understanding what a survivor's life will look like.

5.2 The role of text generation for personalized treatment information

Obviously, when we move from generic to personalized information about treatment options, this implies that static information presentation (on a leaflet, say) is no longer feasible. Fortunately, tools for automatic information presentation based on data have been under development for many years (e.g., [50]). Currently, there are two broad classes of systems [24]: rule-based and statistical ones, both existing in many variants. Of the latter category, generative large language models (or LLMs) are the most prominent manifestation, including, for example, OpenAI's ChatGPT system which took the world's fancy late 2022. While LLMs are capable of producing text with impressive depth and fluency, they suffer from a number of problems. There are general concerns about, for example, the inherent biases and stereotyping of LLMs, the ecological footprint of training them and the poor quality for low-resource languages, see e.g., [3] for further discussion.

Two specific problems for current purposes are the following: (1) The texts generated by LLMs need not be factually true and correct, for example, saying that a treatment has a relatively low risk of a particular side-effect while in truth the risk is rather high or unknown; (2) They may hallucinate information, that is: generate output that does not correspond to anything in the input, for example discussing a non-existing or dispreferred, outdated treatment.

For these reasons, it is currently assumed that LLMs are not practically useful in the kind of health contexts we consider in this project, and could even be harmful [1]. Some LLM model-developers therefore explicitly rule out such applications. For example, the HuggingFace website for BLOOM (an open-access LLM with 176B parameters; [69]) explicitly lists the biomedical domain as well as "critical automated decisions" as out-of-scope uses of their model. For all these reasons, in this project we relied on more traditional rule-based text generation models, which are capable of generating a wide-variety of texts, following a number of fixed patterns [35].

As a result, it can be guaranteed that generated output is always factually correct and hallucination-free. The price to be paid for this is that the generated output is not as fluently flexible as LLM-generated text. In future research, more work needs to be done on harnessing the strengths of both methods, paving the way for fluid and flexible generated output that can be controlled for factualness and lack of hallucinations.

5.3 Investigating other modalities during personalized risk communication

A critical note on our approach has been its predominant focus on developing communication methods for representing personalized statistical information in the written, non-interactive communication domain (e.g., presenting statistics and probabilities in web-based patient decision aids). Relatively underexplored has been the effect of non-verbal communication of personalized information on patients' responses during spoken risk communication [25]. For instance, when doctors are using clinical prediction models during consultations with their patients, they can non-verbally signal uncertainty around individualized risk estimates by using vocal cues such as a rising, question-like intonation contour or preceded by filled pauses (e.g., "uhm...") [16].

So far, these non-verbal or prosodic cues have largely been ignored by risk communication researchers, while other studies show that they may have a substantially larger impact on people's risk and uncertainty perceptions than the risk estimates themselves [65]. As a next step for future research, we should discover which non-verbal communication strategies of disclosing risk information during patient-doctor interactions are most effective while also mitigating potential negative psychological responses such as fear or losing trust. Understanding these effects will help doctors to communicate individualized outcome information obtained via prediction models to their patients in a transparent and understandable way.

5.4 Communicating to patients with lower health literacy

When dealing with health information provision to patients, it is crucial that this information is understandable for all. However, patients might have lower health literacy skills, come from diverse socio-economic backgrounds and have different education levels. Especially when designing eHealth solutions, it is therefore crucial to take into account the (digital) health literacy of patients. A definition of (digital) health literacy is "the degree to which individuals have the capacity to obtain, process, and understand basic [digital] health information and services needed to make appropriate health decisions" [46, 18]. Patients with lower health literacy skills are less likely to participate in research than patients with higher health literacy skills, better socio-economic background and higher education levels [33, 22]. We recommend researchers to work closely together with organizations that have ample experience with communicating to this specific group and advise to measure health literacy skills so that samples can be described and recruitment strategies may be adjusted when few patients with low health literacy have been recruited. Additionally, we also conducted some studies with a representative sample of non-patients as some effects (e.g., effects on knowledge, utility) can be tested without burdening vulnerable patient populations.

5.5 Lessons learned about the collaborations

Besides involving patients in our project, we collaborated with many other stakeholders. The starting point of our project was the collaboration between Tilburg University and the Netherlands Comprehensive Cancer Organization (IKNL). This collaboration ensured accessibility to data and a mix of researchers from different disciplines (e.g., statistics, communication, data science, epidemiology, linguistics, (clinical) psychology). Although each discipline often had their own terminology, we found that bringing together ideas from different fields strengthened the feasibility of explaining personalized predictions in practice.

The close collaboration with IKNL was unusual, and a key-factor to the success of the project. The collaboration started with the preparation of the project proposal, and continued throughout. We could not have carried out this project without access to and support for using the NCR and the PROFILES registry. The PhD researchers working on this project also spend part of their time at IKNL, facilitating this process. Additionally, monthly consortium meeting helped in making sure that research ideas were grounded in practice, and the feasibility and applicability of the approach were a constant point of attention in these meetings. Finally, IKNL provided a bridge to actual hospital practice, which was instrumental in the evaluation studies we conducted.

During the project we expanded the scope to different health domains in oncology and physical trauma by working closely together with the Antoni van Leeuwenhoek Hospital (skin, bladder, breast cancer; data set: quality of life data obtained through the EORTC-QLQ-C30), the Catharina Hospital (colorectal cancer; data set: PROFILES, NCR) and the Elisabeth

TweeSteden Hospital (physical trauma; data set Brabant Injury Outcome Surveillance). This gave us the rare opportunity to work on a solution that would generalize to different disease domains, and provided insights into different patient communication traditions across health areas, which are not usually documented. Finally, we also collaborated with different information providers (Kanker.nl, ZorgkeuzeLab, Interactive Studios) that are experienced with communicating complex health information in an understandable way to patients.

5.6 Implications for (clinical) practice

The idea that patients should receive personalized information might seem simple, the clinical reality is unruly. Estimates show that it takes around 17 years for research evidence to be implemented in clinical practice [43] and although eHealth solutions might seem promising, uptake can be low [59]. Nevertheless, the findings of our project also have immediate implications for real-world clinical practice, including patient decision aid developers, healthcare professionals, and general websites about cancer such as online cancer communities. For instance, as a result of the findings presented in this paper, a real-life web-based tool "Specifieke Cijfers" (Specific Numbers) will be launched in 2023 on the Dutch online cancer community website Kanker.nl (`https://kanker.nl`), which will be able to communicate personalized cancer statistics such as (conditional) survival to patients and their relatives. In this tool, visitors will have the opportunity to enter a number of personal and disease related characteristics, for which in return they will see the number of similar patients that are still alive 5 years after diagnosis. Similar to the statistical prediction models outlined above, these personalized cancer statistics are derived from a real-life connection with NCR. Moreover, in line with our project's findings, the most recent version of this tool discloses personalized statistical information while keeping the information and data entry characteristics short and concise, and also provides contextual information by explaining the gist (or bottom-line) meaning of conditional survival outcomes. Finally, insights from our project have already been acknowledged, valued, and adopted outside the oncology domain, such as the trauma setting where patients need to be informed about their recovery after an injury. For this, our data-driven approach for determining personalized narratives will soon be developed and implemented in the Patient Journey App, a patient-centered application for patients used by the Elisabeth Tweesteden Hospital in the Netherlands.

References

1 A. Arora and A. Arora. The promise of large language models in health care. *The Lancet*, 401:641, 2023.

2 L. A. Baezconde-Garbanati, J. S. Chatterjee, L. B. Frank, S. T. Murphy, M. B. Moran, L. N. Zhao N. Werth, P. Amezola de Herrera, D. Mayer, J. Kagan, and D. O'Brien. Tamale lesson: A case study of a narrative health communication intervention. *Journal of Communication in Healthcare*, 7(2):82–92, 2014.

3 E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

4 N. Bol, E. Smit, and M. Lustria. Tailored health communication: Opportunities and challenges in the digital era. *Digital Health*, 6:1–3, 2020.

5 C. Bonner, L. J. Trevena, W. Gaissmaier, P. K. J. Han, Y. Okan, E. Ozanne, E. Peters, D. Timmermans, and B. J. Zikmund-Fisher. Current best practice for presenting probabilities in patient decision aids: fundamental principles. *Medical Decision Making*, 41(7):821–833, 2021.

6 E. Boomstra, S. Hommes, R. D. Vromans, A. M. van den Burg, A. M. Schrijver, M. W. J. M. Wouters, I. M. C. van der Ploeg, M. W. van de Kamp, E. J. Krahmer, L. V. van de Poll-Franse, and K. de Ligt. Numbers call for action, personalized narratives provide support and recognition": A qualitative assessment of cancer patients' perspectives on patient reported outcome measures (proms) feedback with narratives. *Journal of Cancer Survivorship*, 2024.

7 C. Charles, A. Gafni, T. Whelan, and M. A. O'Brien. Treatment decision aids: conceptual issues and future directions. *Health Expectations*, 8(2):114–125, 2005.

8 S. S. Chauhan, J. Kaur, M. Kumar, A. Matta, G. Srivastava, A. Alyass, J. Assi, I. Leong, C. MacMillan, I. Witterick, T. J. Colgan, N. K. Shukla, A. Thakar, M. C. Sharma, K. W. M. Siu, P. G. Walfish, and R. Ralhan. Prediction of recurrence-free survival using a protein expression-based risk classifier for head and neck cancer. *Oncogenesis*, 4(4):e147, 2015.

9 P. C. Chen, Y. M. Yeh, B. W. Lin, R. H. Chan, P. F. Su, Y. C. Liu, C. T. Lee, S. H. Chen, and P. C. Lin. A prediction model for tumor recurrence in stage ii-iii colorectal cancer patients: From a machine learning model to genomic profiling. *Biomedicines*, 10(2), 2022.

10 F. J. Clouth, M. J. Bijlsma, S. Pauws, and J. K. Vermunt. The parametric g-formula for latent Markov models, 2024. PsyArXiv preprint.

11 F. J. Clouth, A. Moncada-Torres, G. Geleijnse, F. Mols, F. N. van Erning, I. H. J. T. de Hingh, S. C. Pauws, L. V. van de Poll-Franse, and J. K. Vermunt. Heterogeneity in quality of life of long-term colon cancer survivors: A latent class analysis of the population-based profiles registry. *The Oncologist*, 26(3):e492–e499, 2021.

12 F. J. Clouth, S. Pauws, F. Mols, and J. K. Vermunt. A new three-step method for using inverse propensity weighting with latent class analysis. *Advances in Data Analysis and Classification*, 16(2):351–371, 2022. doi:10.1007/S11634-021-00456-5.

13 F. J. Clouth, S. Pauws, and J. K. Vermunt. Three-step latent class analysis with inverse propensity weighting in the presence of differential item functioning. *Structural Equation Modeling: A Multidisciplinary Journal*, 30(5):737–748, 2023. doi:10.1080/10705511.2022.2161384.

14 A. De Graaf, J. Sanders, and H. Hoeken. Characteristics of narrative interventions and health effects: A review of the content, form, and context of narratives in health-related narrative persuasion research. *Review of Communication Research*, 4:88–131, 2016.

15 B. H. de Rooij, I. Ramsey, F. J. Clouth, N. Corsini, J. S. Heyworth, B. M. Lynch, J. K. Vallance, and T. Boyle. The association of circadian parameters and the clustering of fatigue, depression, and sleep problems in breast cancer survivors: a latent class analysis. *Journal of Cancer Survivorship*, 17(5):1405–1415, 2022.

16 C. Dijkstra, E. Krahmer, and M. Swerts. Manipulating uncertainty: The contribution of different audiovisual prosodic cues to the perception of confidence. In *Proceedings of the Speech Prosody Conference*, pages 1–4, 2006.

17 J. L. Donovan, F. C. Hamdy, J. A. Lane, M. Mason, C. Metcalfe, E. Walsh, ..., and ProtecT Study Group. Patient-reported outcomes after monitoring, surgery, or radiotherapy for prostate cancer. *New England Journal of Medicine*, 375(15):1425–1437, 2016.

18 P. Dunn and E. Hazard. Technology approaches to digital health literacy. *International Journal of Cardiology*, 293:294–296, 2019.

19 G. Elwyn, D. Frosch, R. Thomson, N. Joseph-Williams, A. Lloyd, P. Kinnersley, Cording E, Tomson D, Dodd C, Rollnick S, Edwards A, and M. Barry. Shared decision making: a model for clinical practice. *Journal of General Internal Medicine*, 27:1361–1367, 2012.

20 E. G. Engelhardt, A. H. Pieterse, N. van Duijn-Bakker, J. R. Kroep, H. C. de Haes, E. M. Smets, and A. M. Stiggelbout. Breast cancer specialists' views on and use of risk prediction models in clinical practice: a mixed methods approach. *Acta Oncologica*, 54(3):361–367, 2015.

21 A. Fagerlin, B. J. Zikmund-Fisher, and P. A. Ubel. If i'm better than average, then i'm ok? comparative information influences beliefs about risk and benefits. *Patient Education and Counseling*, 69(1-3):140–144, 2007.

22 J. G. Ford, M. W. Howerton, G. Y. Lai, T. L. Gary, S. Bolen, M. C. Gibbons, J. Tilburt, C. Baffi, T. P. Tanpitukpongse, R. F. Wilson, N. R. Powe, and E. B. Bass. Barriers to recruiting underrepresented populations to cancer clinical trials: A systematic review. *Cancer*, 112:228–242, 2008.

23 S. J. Franssen, S. M. Lagarde, J. R. van Werven, E. M. Smets, K. T. Tran, J. T. M. Plukker, J. J. van Lanschot, and H. C. de Haes. Psychological factors and preferences for communicating prognosis in esophageal cancer patients. *Psycho-Oncology: Journal of the Psychological, Social and Behavioral Dimensions of Cancer*, 18(11):1199–1207, 2009.

24 A. Gatt and E. Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018. `doi:10.1613/JAIR.5477`.

25 J. Geipel, C. Hadjichristidis, L. Savadori, and B. Keysar. Language modality influences risk perception: Innovations read well but sound even better. *Risk Analysis*, 43:558–570, 2022.

26 G. Gigerenzer, W. Gaissmaier, E. Kurz-Milcke, L. M. Schwartz, and S. Woloshin. Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2):53–96, 2008.

27 N. Haase, C. Betsch, and F. Renkewitz. Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication*, 20(8):920–929, 2015.

28 F. C. Hamdy, J. L. Donovan, J. A. Lane, M. Mason, C. Metcalfe, P. Holding, ..., and ProtecT Study Group. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New England Journal of Medicine*, 375(15):1415–1424, 2016.

29 R. J. M. Havermans, F. J. Clouth, K. W. W. Lansink, J. K. Vermunt, M. A. C. de Jongh, and L. de Munter. Prediction of recovery in trauma patients using latent Markov models. *European Journal of Trauma and Emergency Surgery*, 48(3):2059–2080, 2022.

30 B. He, W. Chen, L. Liu, Z. Hou, H. Zhu, H. Cheng, Y. Zhang, S. Zhan, and S. Wang. Prediction models for prognosis of cervical cancer: Systematic review and critical appraisal. *Frontiers in Public Health*, 9:654454, 2021.

31 L. Heilmann. Health and numeracy: the role of numeracy skills in health satisfaction and health-related behaviour. *ZDM Mathematics Education*, 52:407–418, 2020.

32 M. A. Hernan and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health*, 60:578–586, 2006.

33 R. Hernandez-Ramos, A. Aguilera, F. Garcia, J. Miramontes-Gomez, L. E. Pathak, C. A. Figueroa, and C. R. Lyles. Conducting internet-based visits for onboarding populations with limited digital literacy to an mhealth intervention: Development of a patient-centered approach. *JMIR Formative Research*, 5(4):25299, 2021.

34 C. Hitchens. *Mortality*. Hachette, UK, 2012.

35 S. Hommes, C. van der Lee, F. J. Clouth, X. Verbeek, and E. J. Krahmer. A personalized data-to-text support tool for cancer patients. In *Proceedings of the 12th International Conference on Natural Language Generation, Tokio, Japan*, pages 443–452, 2019.

36 S. Hommes, R. Vromans, M. de Jongh, T. Houwen, N. Bol, E. de Groot, M. Antheunis, and E. Krahmer. Knowing what life has in store for me is better than being in the dark: Trauma patients' perspectives on receiving personalized predictions for life after injury, 2023.

37 S. Hommes, R. D. Vromans, J. Bloemen, S. van Gent, G. Schipper, X. Verbeek, and E. J. Krahmer. Personalized predictions about quality of life after chemotherapy: Development of a data-driven support tool and initial evaluation of colorectal cancer patients and clinicians, 2023.

38 S. Hommes, R. D. Vromans, F. J. Clouth, X. Verbeek, I. de Hingh, and E. J. Krahmer. Communication in decision aids for stage i-iii colorectal cancer patients: A systematic review. *BMJ Open*, 11:e044472, 2021.

39 A. Kalch, C. Küchler, A. Albani, H. Bilandzic, S. Fischer, and I. Kirchberger. On the need for narratives in patient information: Differentiating types and functions of narratives from pulmonary embolism patients' point of view. *Frontiers in Communication*, 7:990819, 2022.

40 K. Kenzik, M. Y. Martin, M. N. Fouad, and M. Pisu. Health-related quality of life in lung cancer survivors: Latent class and latent transition analysis. *Cancer*, 121(9):1520–1528, 2015.

41 M. W. Kreuter, M. Green, J. Cappella, M. Slater, M. Wise, D. Storey, E. M. Clark, D. J. O'Keefe, D. O. Erwin, K. Holmes, L. J. Hinyard, T. Houston, and S. Woolley. Narrative communication in cancer prevention and control: A framework to guide research and application. *Annals of Behavioral Medicine*, 33(3):221–235, 2007.

42 L. E. Kunst, M. A. L. M. van Assen, F. J. Clouth, C. Hunt, M. Abbott, J. Maas, and M. H. J. Bekker. Who cares what other people think? a longitudinal investigation of long-term self-esteem stability and its associations with autonomy-connectedness, 2024. Manuscript submitted for publication.

43 Z. S. Morris, S. Wooding, and J. Grant. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, 104(12):510–520, 2011.

44 Netherlands Cancer Registry (ncr) and Netherlands Comprehensive Cancer Organisation. (iknl), derived via www.iknl.nl/en/ncr/ncr-data-figures [03/31/2023].

45 S. M. Noar, C. N. Benac, and M. S. Harris. Does tailoring matter? meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin*, 133(4):673–693, 2007.

46 R. Parker and S. C. Ratzan. Health literacy: A second decade of distinction for americans. *Journal of Health Communication*, 15(2):20–33, 2010.

47 S. Pauws, A. Gatt, E. Krahmer, and E. Reiter. Making effective use of healthcare data using data-to-text technology. In *Data Science for Healthcare*, pages 119–145. Springer, 2019. doi:10.1007/978-3-030-05249-2_4.

48 E. Peters. *Innumeracy in the wild: Misunderstanding and misusing numbers*. Oxford University Press, 2020.

49 PROFILES. derived via. URL: https://www.dataarchive.profilesregistry.nl/publications.

50 E. Reiter and R. Dale. *Building Natural Language Generation Systems*. Cambridge University Press, 2000.

51 S. Salem, T. Weskott, and A. Holler. Does narrative perspective influence readers' perspective-taking? an empirical study on free indirect discourse, psycho-narration and first-person narration. *Glossa: A Journal of General Linguistics*, 2(1):61, 2017.

52 V. A. Shaffer, E. S. Focella, A. Hathaway, L. D. Scherer, and B. J. Zikmund-Fisher. On the usefulness of narratives: An interdisciplinary review and theoretical model. *Annals of Behavioral Medicine*, 52(5):429–442, 2018.

53 V. A. Shaffer and B. Zikmund-Fisher. All stories are not alike: A purpose-, content- and valance-based taxonomy of patient narratives in decision aids. *Medical Decision Making*, 33(1):4–13, 2013.

54 F. Shen, V. C. Sheer, and R. Li. Impact of narratives on persuasion in health communication: A meta-analysis. *Journal of Advertising*, 44(2):105–113, 2015.

55 R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics. *2016. CA: A Cancer Journal for Clinicians*, 66(1):7–30, 2016.

56 L. V. van de Poll-Franse, N. Horevoorts, M. van Eenbergen, J. Denollet, J. A. Roukema, N. K. Aaronson, A. Vingerhoets, J. W. Coebergh, J. de Vries, M. L. Essink-Bot, and F. Mols. The patient reported outcomes following initial treatment and long term evaluation of survivorship registry: scope, rationale and design of an infrastructure for the study of physical and psychosocial outcomes in cancer survivorship cohorts. *European Journal of Cancer*, 47(14):2188–2194, 2011.

57 H. G. van den Boorn, E. G. Engelhardt, J. van Kleef, M. A. G. Sprangers, M. G. H. van Oijen, A. Abu-Hanna, A. H. Zwinderman, V. M. H. Coupe, and H. W. M. van Laarhoven. Prediction models for patients with esophageal or gastric cancer: A systematic review and meta-analysis. *PLoS One*, 13(2):e0192310, 2018.

58 N. C. van der Velden, H. W. van Laarhoven, S. A. Burgers, L. E. Hendriks, F. Y. de Vos, A. M. C. Dingemans, J. Jansen, J. W. van Haarst, J. Dits, E. M. Smets, and I. Henselmans. Characteristics of patients with advanced cancer preferring not to know prognosis: A multicenter survey study. *BMC Cancer*, 22(1):1–12, 2022.

59 J. E. van Gemert-Pijnen, N. Nijland, M. van Limburg, H. C. Ossebaard, S. M. Kelders, G. Eysenbach, and E. R. Seydel. A holistic framework to improve the uptake and impact of ehealth technologies. *Journal of Medical Internet Research*, 13(4):e111, 2011.

60 R. Vromans, K. Tenfelde, S. Pauws, M. Van Eenbergen, I. Mares-Engelberts, G. Velikova, L. van de Poll-Franse, and E. Krahmer. Assessing the quality and communicative aspects of patient decision aids for early-stage breast cancer treatment: a systematic review. *Breast Cancer Research and Treatment*, 178:1–15, 2019.

61 R. D. Vromans, S. Hommes, F. J. Clouth, D. N. N. Lo-Fo-Wong, X. A. A. M. Verbeek, L. V. van de Poll-Franse, S. C. Pauws, and E. J. Krahmer. Need for numbers: Assessing cancer survivors' needs for personalized and generic statistical information. *BMC Medical Informatics and Decision Making*, 22(1):1–14, 2022.

62 R. D. Vromans, S. C. Pauws, N. Bol, L. V. van de Poll-Franse, and E. J. Krahmer. Communicating tailored risk information of cancer treatment side effects: Only words or also numbers? *BMC Medical Informatics and Decision Making*, 20:1–12, 2020.

63 R. D. Vromans, S. C. Pauws, L. V. van de Poll-Franse, and E. J. Krahmer. Effects of comparative information when communicating personalized risks of treatment outcomes: An experimental study. *Journal of Risk Research*, 26(3):324–343, 2022.

64 R. D. Vromans, C. N. Tillier, S. C. Pauws, H. G. van der Poel, L. V. van de Poll-Franse, and E. J. Krahmer. Communication, perception, and use of personalized side-effect risks in prostate cancer treatment decision-making: An observational and interview study. *Patient Education and Counseling*, 105(8):2731–2739, 2022.

65 R. D. Vromans, C. C. M. van de Ven, S. J. W. Willems, E. J. Krahmer, and M. Swerts. It is, uh, very likely? the impact of prosodic uncertainty cues on the perception and interpretation of spoken verbal probability phrases. *Risk Analysis*, pages 1–20, 2024.

66 R. D. Vromans, M. C. van Eenbergen, G. Geleijnse, S. C. Pauws, L. V. van de Poll-Franse, and E. J. Krahmer. Exploring cancer survivor needs and preferences for communicating personalized cancer statistics from registry data: Qualitative multimethod study. *JMIR Cancer*, 7(4):e25659, 2021.

67 R. D. Vromans, M. C. van Eenbergen, S. C. Pauws, G. Geleijnse, H. G. van der Poel, L. V. van de Poll-Franse, and E. J. Krahmer. Communicative aspects of decision aids for localized prostate cancer treatment–a systematic review. *Urologic Oncology: Seminars and Original Investigations*, 37(7):409–429, 2019.

68 G. C. Wishart, E. M. Azzato, D. C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, and P. D. P. Pharoah. Predict: a new uk prognostic model that predicts survival following surgery for invasive breast cancer. *Breast Cancer Research*, 12(1):1–10, 2010.

69 BigScience Workshop:, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, ..., and P. Villegas. Bloom: A 176b-parameter open-access multilingual language model, 2022. arXiv preprint. `arXiv:2211.05100`.

70 M. Zeguers, H. C. de Haes, L. C. Zandbelt, C. L. Ter Hoeven, S. J. Franssen, D. D. Geijsen, C. C. Koning, and E. M. Smets. The information needs of new radiotherapy patients: how to measure? do they want to know everything? and if not, why? *International Journal of Radiation Oncology, Biology, Physics*, 82(1):418–424, 2012.

71 B. J. Zikmund-Fisher. The right tool is what they need, not what we have: a taxonomy of appropriate levels of precision in patient risk communication. *Medical Care Research and Review*, 70(1), 2013.

72 B. J. Zikmund-Fisher. Helping people know whether measurements have good or bad implications: increasing the evaluability of health and science data communications. *Policy Insights from the Behavioral and Brain Sciences*, 6(1):29–37, 2019.

SeNiors empOWered via Big Data to Joint-Manage Their Medication-Related Risk of Falling in Primary Care: The SNOWDROP Project

Leonie Westerbeek[1]
Amsterdam School of Communication Research/ASCoR, University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Noman Dormosh
Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

André Blom
Uw Zorg Online, Amsterdam, The Netherlands

Martijn Heymans
Department of Epidemiology and Data Science, Amsterdam UMC location Vrije Universiteit Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Meefa Hogenes
Department of Medical Content, ExpertDoc, Rotterdam, The Netherlands

Annemiek Linn
Amsterdam School of Communication Research/ASCoR, University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Stephanie Medlock
Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Martijn Schut
Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, The Netherlands
Department of Laboratory Medicine, Amsterdam UMC location Vrije Universiteit Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Nathalie van der Velde
Department of Internal Medicine, Section of Geriatric Medicine, Amsterdam UMC location University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Henk van Weert
Department of General Practice, Amsterdam UMC location University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Julia van Weert
Amsterdam School of Communication Research/ASCoR, University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Ameen Abu-Hanna
Department of Medical Informatics, Amsterdam UMC location University of Amsterdam, The Netherlands
Amsterdam Public Health Research Institute, The Netherlands

Abstract

In older persons, falls are the leading cause of injuries, often resulting in emergency room visits, serious injuries, and possibly even death. Medications are a major risk factor for falls. Because we lack tools to assess individualized risks, general practitioners (GPs) struggle with fall related

[1] corresponding author

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 4; pp. 4:1–4:12
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

medication management for seniors, and senior patients are not properly equipped to engage in the joint management of their medications. Our aim in this project is to develop and evaluate a comprehensive data-driven science approach for valid prediction of personalized risk of falling that effectively supports joint medication management between seniors and GPs. The project has two objectives. First, we aim to develop and validate prediction models from electronic health records for assessing individualized risk of medication-related falls. Data science challenges include free text analysis; accounting for missing values; searching medication hierarchies; engineering new predictors, and understanding limitations of our approach. Second, we aim to develop and evaluate a joint medication management strategy for older patients and GPs, consisting of a clinical decision support system (CDSS) and a patient portal. We evaluate the effects of this strategy on changes in the quality of shared decision-making during a medication review consultation, medication management, and patient outcomes. The learnings from this project and the architecture underpinned by predictive modelling to support both GPs and patients can also be applied to other major health problems in the future.

2012 ACM Subject Classification Applied computing → Life and medical sciences

Keywords and phrases accidental falls, fall risk, medication, prediction model, clinical decision support, patient portal, shared decision-making

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.4

1 Introduction

An important ongoing transformation in public health is a shift from a disease-centered to a patient-centered approach, and from single-disease-focused to multi-disease-focused care. This transformation requires high quality shared decision-making (SDM) between patients and clinicians. During the SDM process healthcare related decisions are made by the patient and healthcare professional together, exploring the best available treatment options and the patient's goals [11]. SDM is particularly important in the context of medication management among older patients. Older people use a relatively large number of medicines and are sensitive to side effects. Falls are a prevalent side effect and are the leading cause of traumatic injury among older people [22]. Just in the Netherlands in 2022, every 4 minutes an older person (65 years and older) was admitted to the emergency room due to a fall, 87,000 older people were seriously injured (e.g. brain damage, hip fractures), and 5995 older people died because of a fall [21]. Furthermore, falls can set off a downward spiral, where fear of falling can trigger reduced activity, and hence impaired balance, and decreased strength, ultimately leading to a lower quality of life [19].

Certain medications are recognized as an important -but modifiable- predisposing factor for falls (also called Fall Risk Increasing Drugs; FRIDs), providing an opportunity to minimize the fall risk when appropriately managed [12, 14]. It is therefore important that general practitioners (GPs) and older patients regularly evaluate medication use together. However, GPs lack tools to assess individualized risks, and struggle with fall-related medication management for older patients. Older patients, in turn, are often not properly equipped to engage in SDM to reduce their fall-risk. An explanation for this might be the lack of knowledge; about 85% of older patients with polypharmacy do not know the indication of their prescribed medication [2]. Polypharmacy entails taking multiple medications (e.g. five or more) simultaneously [13]. In this interdisciplinary project, called SNOWDROP (**S**e**N**iors emp**OW**red via big **D**ata to joint-manage their medication-related **R**isk **O**f falling in **P**rimary care), we aim to achieve two objectives: to develop and validate prediction models for falls in older patients that can be used to estimate individualized fall risk (WP1; data science), and

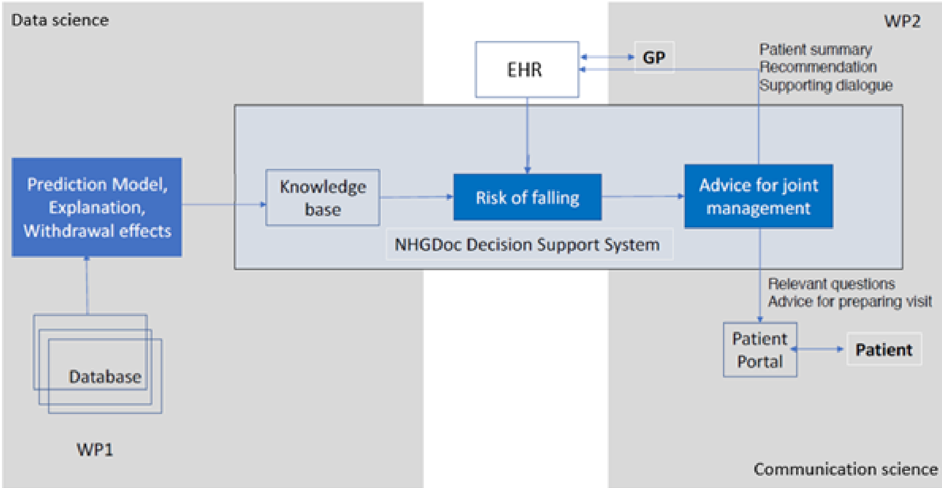


Figure 1 Visualized overview of data science (WP1) and communication science (WP2).

to use these prediction models to provide smart decision support for shared decision making between GPs and older patients (WP2; communication science). A visual representation of both parts can be found in Figure 1.

SNOWDROP spans a wide range of research and development activities relevant to medical-, data-, and communication sciences. In the data science part, we address the analysis of routinely collected "messy" data in the electronic health record (EHR) of the GPs, mining clinical notes, handling missing values, searching for potential predictive variables, and ultimately using statistical machine learning techniques to develop and validate prediction models for falls.

In the communication science part of the project (WP2), based on the prediction models obtained in WP1, personalized decision support about the risks of individual patients' medication will be provided to the GP via a clinical decision support system (CDSS). Patients, in turn, will be supported through a patient portal to prepare for the medication review consultation that they will be invited to. In WP2, both systems are systematically developed according to user-centered design (UCD) principles. UCD entails involving end users in all phases of the development process, to make sure the end product fits their needs and wishes seamlessly [3, 16]. We investigate the effects of the intervention on, for instance, SDM in a randomized controlled trial (RCT) in which older patients and GPs in the intervention group will receive the full intervention consisting of both systems (the CDSS and the patient portal), while patients in the control group will receive usual care. The health communication researchers are experts on, for instance, doctor-patient communication (including the SDM component of the SNOWDROP intervention), motivating patients and healthcare professionals to use the intervention, and persuasive health technologies, including the use of UCD methods to ensure that the intervention fits the needs and wishes of the users in order to increase acceptance/compliance.

2 Methods

The first objective of SNOWDROP is to develop and validate a prediction model for falls in patients aged 65 or older using EHR data collected from GPs in the Netherlands. The EHR data comprise demographics, diagnoses, prescribed medications and clinical notes written mainly by GPs. A prediction model is a mathematical formula that connects various predictors related to a specific individual to estimate the likelihood of a particular outcome (falls in our case) in the future [15]. It is important to note that predictive models are designed to estimate the likelihood of an event based on specified variables. Unlike causal inference, which was not the primary objective here, predictive modeling does not necessarily require adjusting for confounding factors. Causal inference, on the other hand, specifically involves understanding how the likelihood of an event changes when we intervene by altering a predictor, often necessitating careful adjustment for confounding factors to ensure accurate conclusions. Furthermore, the prediction model in this project is solely used to identify patients at risk of falling. The recommendations that are subsequently provided by the CDSS are based on guidelines.

Potential predictors (i.e., input for the prediction model) for falls were identified according to the literature and expert knowledge. Lab measurements were excluded from the analysis due to a high percentage of individuals lacking measured values (46%-100%). Attempting fall prediction with imputed values did not enhance performance. Using the XGBoost machine learning algorithm [4], which accommodates missing values, did not render these variables predictive. We used two independent cohorts, namely the development cohort and the validation cohort, to develop and assess the generalizability of our prediction model. The development cohort contained data on 36,470 older patients enlisted with 50 GPs, and the validation cohort contained data on 39,342 older patients enlisted with 59 other GPs. The validation step is necessary to justify the implementation of the prediction model in practice. The advantage of using a prediction model using EHR data is that it is based on variables that are routinely collected during medical care. As such, our prediction model can be integrated with a CDSS in an EHR system to identify patients at risk for which the CDSS can then provide (guideline-based) advice to manage FRIDs in order to decrease the fall risk. The specification of the knowledge base which underlies the pieces of advice to manage FRIDs was made in collaboration with the ADFICE_IT study [7].

We further applied Natural Language Processing (NLP) techniques to evaluate the incremental predictive value of the unstructured data (i.e., clinical notes) over the structured data (i.e., age, sex and medication). NLP entails applying computational techniques for automated analysis of texts and spoken words [5]. Our approach relies on modern NLP techniques that translate words and documents into a vector representation, in order to capture their meaning. Specifically, we applied topic modelling [6, 17] using Top2Vec [arXiv:2008.09470] to extract latent topics from the clinical notes. The resulting topics were used as input variables to a machine learning algorithm to predict future falls.

The second objective is to design and evaluate a SDM strategy for GPs and older patients. The prediction model, together with the advice, is embedded in NHGDoc, an existing decision support system, that provides personalized information relevant to GPs via their EHR. NHGdoc can connect to every Dutch GP information system except for one. This makes future implementation of the intervention easily feasible for the majority of Dutch systems. Information relevant to patients is provided via an existing Dutch patient portal called Uw Zorg Online. We evaluate the usability and the effects of this strategy on changes in the quality of the communication, in particular SDM, during a fall-related medication review consultation, medication management, and patient outcomes (e.g., beliefs about medicines, recall, and decisional conflict).

To make sure that the intervention is evidence based and matches the needs and wishes of its end users (i.e. GPs and older patients) the Medical Research Council (MRC) guideline for complex interventions is followed [18]. This guideline consists of four phases; (1) development,

(2) feasibility, (3) implementation, and (4) evaluation. In order to develop the intervention, a UCD approach is applied, meaning that end users are closely involved during all phases of the development process. During the development and feasibility phases several UCD methods were applied. During the development phase, a systematic review concerning barriers and facilitators for using a medication-related CDSS as perceived by clinicians was conducted [24]. Furthermore, interviews with older patients (N = 12 patients) and focus groups with GPs (N = 13 GPs) were performed to explore their needs and wishes for both systems and the intervention [23]. Based on these three studies, the first prototypes of the patient portal and CDSS were developed in collaboration with the partners Uw Zorg Online (formerly known as Pharmeon) and ExpertDoc respectively (see section partners for more information on the collaborations).

In the second phase -feasibility- the first prototypes of both systems were tested. This was done by conducting individual usability tests with GPs (N = 5) to test the CDSS, and individual usability tests with older patients (N = 5) to test the patient portal, both using a think-aloud method. Participants (both GPs and older patients) were presented with hypothetical cases asking them to perform certain tasks while using the system. The think-aloud method that we used entails that while carrying out the tasks, GPs and patients were prompted to think aloud and verbalize their thought processes. Afterwards, participants were interviewed to gather additional feedback on the systems. Based on these usability tests, the prototypes were improved. This led to the third phase, implementation. Our partners ExpertDoc and Uw Zorg Online implemented the improvements and used this to develop ready-to-implement versions of both systems. GPs were trained on how to use the system and the intervention as a whole by means of a web lecture and an in-person training.

The fourth and final phase -evaluation- consists of an RCT in which the complete intervention is tested. Six Dutch primary care practices participate in this RCT to test the SNOWDROP intervention. General practices were randomized to either the intervention condition or the control condition. Each participating GP is asked to perform +/- 14 consultations with older patients with a high fall risk to discuss their fall-related medication. GPs in the intervention condition receive the full intervention. The patients of the participating GPs in this condition prepare their visit using the patient portal. GPs in the control condition are also asked to perform a medication review focused on the fall risk, but only receive a simple list of FRIDs. Consultations are audio recorded and analysed for SDM. Patients fill out questionnaires prior to, and after the consultation to assess secondary outcomes. These are beliefs about medicines, recall, and decisional conflict, technology acceptance, and website satisfaction. To check that the systems are actually being used during the RCT, we ask patients which components of the portal they used. For GPs, the audio recordings are checked for evidence that the system was used. Results of the RCT are expected in 2024.

3 Key Results

With respect to the data science part of the project (WP1), we developed a prediction model for falls in older people using primary care EHR data [10]. The model was developed using Bootstrap-enhanced penalized logistic regression with the least absolute shrinkage and selection operator [1, 20], and the prediction strategy was internally validated using 10-fold cross-validation. The model comprises ten predictors that were predictive for falls. These predictors are a combination of demographics, medication, and clinical conditions. The discriminative ability of our model as measured by the area under the receiver operating characteristic curve (AUC) was 0.705. This performance compares favorably and often surpasses performance of previously published models. Furthermore, our model performed

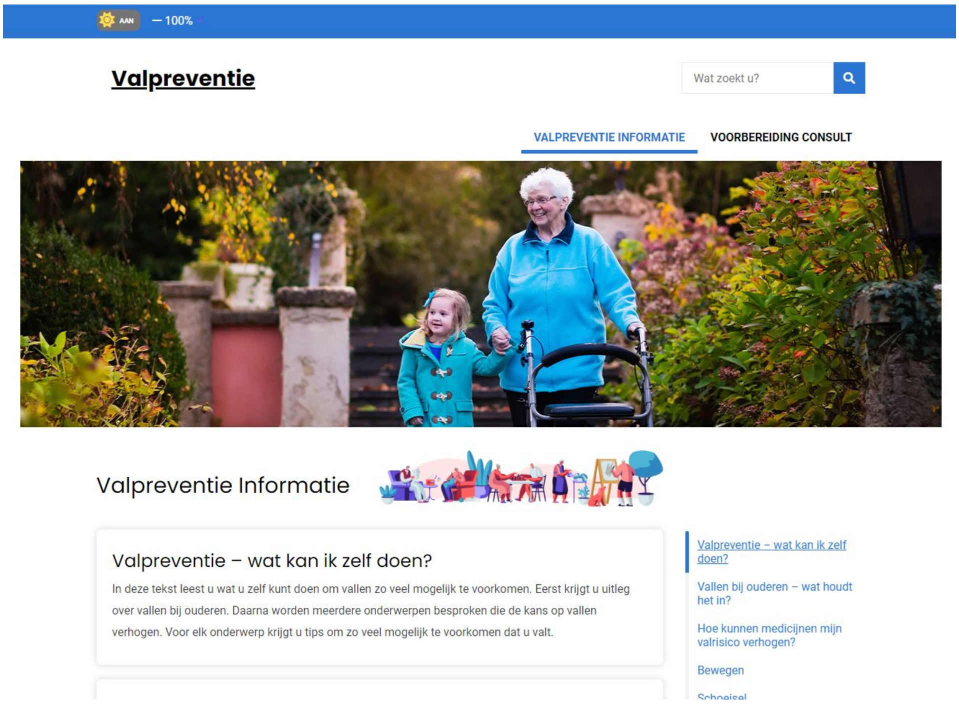


Figure 2 Patient portal.

well when externally validated on an independent cohort sample (AUC in the validation was 0.690), rendering it ready to be trialled in practice [8]. Furthermore, we explored the predictive performance of topics that were automatically extracted from the clinical free-text notes written by GPs to predict falls, and their incremental predictive value over the clinical variables [9]. We used modern NLP techniques that translate words and documents into a vector representation, to capture their meaning and extract the topics. The AUC of the model which incorporated both topics extracted from the clinical notes and clinical variables was 0.718 (CI 95% 0.708-0.727), and the AUC of the model that used only clinical variables was 0.709 (CI 95% 0.700–0.719). We found that the clinical notes form an additional viable data source to develop and improve prediction models for falls compared to traditional prediction models. The strengths of the abovementioned studies includes the use of a large multicentre sample of older people to develop the model, conducting external validation which is rarely performed to assess the generalizability of the prediction models, and leveraging NLP and machine learning to extract information from the clinical notes to predict future falls.

For WP2, several key results have been obtained so far. The systematic review on barriers and facilitators for using a medication-related CDSS as reported by clinicians resulted in a very broad overview of barriers and facilitators. In total, 327 barriers and 291 facilitators were identified [24]. Salient barriers or facilitators (i.e. mentioned in more than one study) were aggregated, resulting in 195 unique barriers and 174 unique facilitators. These were categorized within the Human, Organization and Technology-fit (HOT-fit) model. The results show that the most often reported (and presumably most important) barriers and facilitators were related to (a lack of) usefulness and relevance of information, and ease of use and efficiency of the system. We incorporated the findings of the systematic review throughout the design process; for example, we had a GP evaluate the formulation of every piece of advice to ensure that it was useful and relevant. Usability testing helped us to make sure the system was considered easy to use.

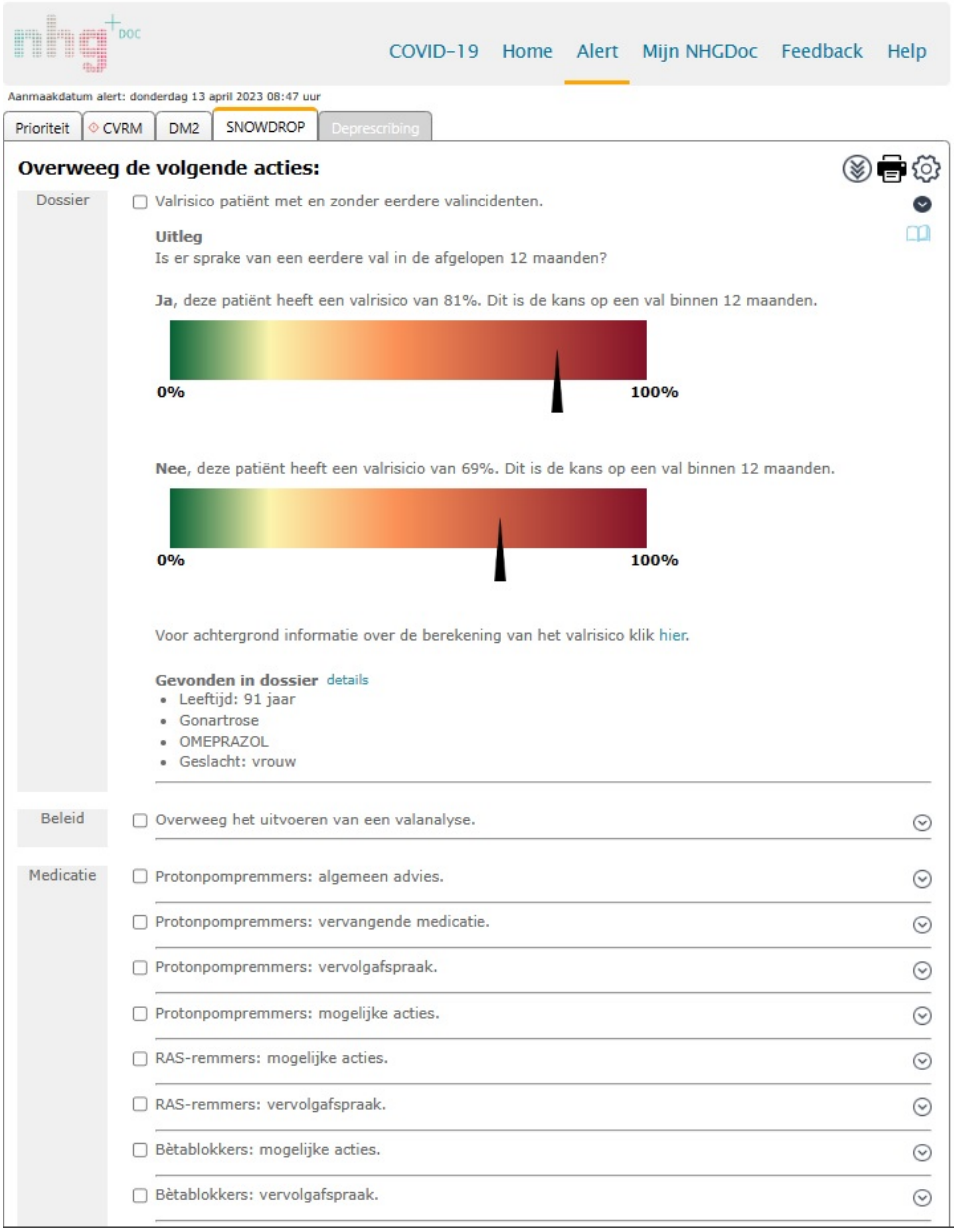


Figure 3 CDSS.

The interviews with patients and focus groups with GPs [23] resulted in the first prototypes of the patient portal and CDSS. These were subsequently improved based on usability testing with a think-aloud approach and interview questions. Results of these studies informed the development and design decisions made in these systems. For instance, a visual presentation of the personalized fall risk, in the form of a gradient scale ranging from bright green to dark red, was incorporated within the CDSS based on the preferences indicated by GPs during the focus groups. Within the patient portal, a question prompt list was implemented. This allows the patient to prepare for the consultation by indicating which questions and/or concerns they would like to discuss with the GP. Figure 2 displays the final, ready-to-implement patient portal and figure 3 displays the final, ready-to-implement CDSS, both which are currently being tested in the RCT. Results of the RCT are expected at end of 2023.

4 Discussion

4.1 Partners

To achieve the abovementioned objectives, we collaborate with three private partners: Elsevier, ExpertDoc and Uw Zorg Online. The motivation of these private partners to participate in this study stems from their vision of the future of healthcare and the role that they would play in it. Investigating and shaping new models for delivering healthcare and the role that technology plays aligns perfectly with our research agenda. First, Elsevier is proactively searching for new “publishing” opportunities in terms of disseminating knowledge. As a publisher, they would like to explore the move from “read this” to data-driven, patient-specific, timely “do this” recommendations to disseminate clinical knowledge into practice. Elsevier is partnering with many large EHR vendors, providing them with clinical decision support plug-ins at the point of care. Elsevier has the clinical knowledge base and world-wide scale to disseminate the insights and systems developed during this project to practicing physicians. They are already focusing heavily on artificial intelligence, machine learning, natural language processing (NLP) and have built a competitive IT division of over 1000 technologists.

Second, ExpertDoc developed an existing CDSS called NHGDoc. This system serves thousands of GPs in the Netherlands. Their product currently consists of clinical decision rules that are solely based on primary care guidelines to provide input (e.g. alerts and reminders) to GPs. ExpertDoc recognizes the opportunity in the increasing availability of digital data, algorithms, and computational infra-structure and the value of these in providing new data-based services to their clients. In this way, machine learning can potentially enhance evidence-based practice with data-driven statistics.

Last, Uw Zorg Online provides two main services to GPs: developing and maintaining websites for their practices, and facilitating patient portals. GPs and other healthcare providers offer a wide range of online services to patients though these systems. Uw Zorg Online is seeking ways to increase the value of their portals to both patients and GPs, hence their desire to collaborate. By answering scientific questions, the academic partners of the consortium will develop the knowledge and the technology in terms of the predictive models and the shared decision-making strategy to fulfil the needs of the private partners.

Collaborating with these parties is both valuable and challenging and resulted in some lessons learned that might be useful for future comparable projects. First of all, it is important to note that collaborating with external technological partners has been crucial for the successful development of our systems. Using their existing infrastructure allowed us to quickly implement a fully integrated CDSS, for instance. This sped up the development process, and also makes possible future implementation in practice more realistic and efficient. However, there are of course also certain limitations that need to be taken into account. Not all of our original plans could realistically be implemented, so it is important to start the process with an open mind and a lot of flexibility. Also later on in the development process, certain issues that were raised concerning the prototypes could not be (fully) resolved because of the existing software that we were working with. Lastly, after developing and evaluating the intervention, it is very valuable to already be in contact with these partners. We are currently conducting a feasibility study with multiple relevant stakeholders (e.g. our technological partners, insurance companies, GPs) to create a blueprint for future implementation of the intervention.

4.2 Challenges

The broad, interdisciplinary nature of this project is very valuable, but also comes with certain challenges. Originally, we planned a seamless integration between the CDSS and the patient portal, meaning that these two systems would be able to communicate and information inserted by the patient would be sent directly to the GP's system. This integration was part of a larger national project which was postponed. This necessitated thinking of other solutions where we could still test the intervention without this strong integration. In the end, we opted for a simpler architecture: during the RCT, the patient is not invited though Uw Zorg Online's original patient portal, but though a new website that Uw Zorg Online designed specifically for the SNOWDROP project.

Recruiting older patients and especially GPs during, and right after the COVID19 pandemic, a time during which there is a large primary care backlog, also posed a challenge. GPs felt, and often still feel, overwhelmed and unable to participate in research. Even though four GPs expressed their intention to participate at the start of the project, in the end they were not able to. For the RCT, fortunately, not all GPs have to start at the same time, allowing us to recruit in a phased manner. The simpler architecture of the patient portal discussed above also presented an opportunity. Now that Uw Zorg Online built a separate website instead of using their original patient portal, practices that don't make use of the Uw Zorg Online patient portal are also eligible for the RCT. This allowed us to recruit from a larger pool of GPs.

From the technical perspective, our project involved the use of textual data (clinical notes) mainly to define the outcome "fall" as the coding system used by the GPs to encode diseases and symptoms does not contain a code to describe falls. Because of the necessity to inspect free-text data, there were inconvenient regulatory procedures in place to ensure privacy protection before acquiring the data. Furthermore, the development of our prediction model involved a manual chart review of the clinical notes to identify fallers, which was laborious and time-consuming. Another technical challenge was the need for a high performance computing environment to combine and analyze the data, especially to apply advanced deep learning algorithms and NLP techniques. Aside from computational demands, there is the important challenge of protecting the privacy of the patients. In this regard, exploring federated learning, where computations occur locally, and only their results are communicated [25], presents a promising avenue for future research.

4.3 Further steps

Although we have achieved key milestones, there are many opportunities for further research. As for the data science aspect (WP1), we will continue leveraging our data, machine learning and NLP to contribute to research and innovations related to falls in older people. Because falls are multifactorial and result from dynamic interaction of risk factors, we will apply data-driven approaches combined with clinical experience to discover these interactions. By doing this, we hope to improve the predictive performance of our prediction model. Furthermore, the temporal ordering of the clinical notes in our dataset is a potential property that can be exploited in different ways. We will apply dynamic topic modelling to discover and isolate topics associated with falls as these topics evolve over time. The hypothesis is that certain topics are more represented in older people who fall and the prominence of these topics increases over time just before fall events. This will improve our understanding of the mechanism underlying falls and could discover potential predictors thereof.

In WP2, the ready-to-implement systems have been developed and are ready to be used in the RCT. This RCT is currently running, meaning that data collection has commenced, but results are not yet available. The next step is to analyze the data and see to what extent the intervention had an effect on our outcome measures. To investigate which steps have to be undertaken for real-life implementation, an additional feasibility study will be conducted. This feasibility study looks at the requirements for broader implementation of the intervention. We will examine how large-scale implementation of the intervention, first in Amsterdam and then nationally, can be achieved. Several stakeholders, such as healthcare professionals and a private partner will be consulted during an interview study. The feasibility study will result in an implementation plan that can be used as a blueprint for future implementation.

5 Conclusion

All in all, the SNOWDROP project has achieved many key milestones and continues to do so. The prediction model has been created and validated, and the full intervention consisting of a CDSS and patient portal has been developed and is currently being evaluated in a RCT. The findings of the RCT will determine future steps of the SNOWDROP project. The learnings from this project as a whole and the architecture underpinned by predictive modelling to support both GPs and patients can also be applied to other major health problems in the future.

References

1 F.R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40, 2008. doi:10.1145/1390156.1390161.

2 D. Bosch-Lenders, D.W.H.A. Maessen, H.E.J.H. Stoffers, J.A. Knottnerus, B. Winkens, and M. Akker. Wat weten ouderen met polyfarmacie van hun pillen? *Nederlands Tijdschrift Voor Geneeskunde*, 160:736, 2016. URL: https://www.ntvg.nl/artikelen/wat-weten-ouderen-met-polyfarmacie-van-hun-pillen.

3 J. Brunner, E. Chuang, C. Goldzweig, C.L. Cain, C. Sugar, and E.M. Yano. User-centered design to improve clinical decision support in primary care. *International journal of medical informatics*, 104:56–64, 2017. doi:10.1016/j.ijmedinf.2017.05.004.

4 C.I. Chen, C.T. Liu, CI Chen, Y.C. Li, and C.C. Chao. Medical errors in a hospital in taiwan: incidence, aetiology and proposed solutions. *J Inf Technol Healthcare*, 2:11–18, 2004. doi:10.1145/2939672.2939785.

5 K.R. Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020. doi:10.1007/978-81-322-3972-7_19.

6 R. Churchill and L. Singh. The evolution of topic modeling. *ACM Computing Surveys*, 54(10s):1–35, 2022. doi:10.1145/3507900.

7 K.K. de Wildt, B. van de Loo, A.J. Linn, S.K. Medlock, S.S. Groos, K.J. Ploegmakers, L.J. Seppala, J.E. Bosmans, A. Abu-Hanna, J.C.M. van Weert, et al. Effects of a clinical decision support system and patient portal for preventing medication-related falls in older fallers: Protocol of a cluster randomized controlled trial with embedded process and economic evaluations (ad f ice_it). *PLoS one*, 18(9):e0289385, 2023. doi:10.1371/journal.pone.0289385.

8 N. Dormosh, M.W. Heymans, N. Velde, J. Hugtenburg, O. Maarsingh, P. Slottje, A. Abu-Hanna, and M.C. Schut. External validation of a prediction model for falls in older people based on electronic health records in primary care. *Journal of the American Medical Directors Association*, 23(10):1691–1697 3, 2022. doi:10.1016/j.jamda.2022.07.002.

9 N. Dormosh, M.C. Schut, M.W. Heymans, O. Maarsingh, J. Bouman, N. Velde, and A. Abu-Hanna. Predicting future falls in older people using natural language processing of general practitioners' clinical notes. *Age and Ageing*, 52(4):1–11, 2023. doi:10.1093/AGEING/AFAD046.

10 N. Dormosh, M.C. Schut, M.W. Heymans, N. Velde, and A. Abu-Hanna. Development and internal validation of a risk prediction model for falls among older people using primary care electronic health records. *The Journals of Gerontology: Series A*, 77(7):1438–1445, 2022. doi:10.1093/GERONA/GLAB311.

11 G. Elwyn, M.A. Durand, J. Song, J. Aarts, P.J. Barr, Z. Berger, N. Cochran, D. Frosch, D. Galasiński, P. Gulbrandsen, P.K.J. Han, M. Härter, P. Kinnersley, A. Lloyd, M. Mishra, L. Perestelo-Perez, I. Scholl, K. Tomori, L. Trevena, and T.Van der Weijden. A three-talk model for shared decision making: multistage consultation process. *BMJ*, 359:4891, 2017. doi:10.1136/BMJ.J4891.

12 A. Ham, M. Swart, A. Enneman, S. Dijk, S. Oliai Araghi, J. Wijngaarden, N. Zwaluw, E. Brouwer-Brolsma, R. Dhonukshe-Rutten, N. Schoor, T. Cammen, P. Lips, C. Groot, A. Uitterlinden, R. Witkamp, B. Stricker, and N. Velde. Medication-related fall incidents in an older, ambulant population: The b-proof study. *Drugs Aging*, 31:917–927, 2014. doi:10.1007/s40266-014-0225-x.

13 N. Masnoon, S. Shakib, L. Kalisch-Ellett, and G.E. Caughey. What is polypharmacy? a systematic review of definitions. *BMC geriatrics*, 17:1–10, 2017. doi:10.1186/s12877-017-0621-2.

14 J. Michalcova, K. Vasut, M. Airaksinen, and K. Bielakova. Inclusion of medication-related fall risk in fall risk assessment tool in geriatric care units. *BMC Geriatrics*, 20(1):454, 2020. doi:10.1186/s12877-020-01845-9.

15 K.G.M. Moons, D.G. Altman, J.B. Reitsma, J.P.A. Ioannidis, P. Macaskill, E.W. Steyerberg, A.J. Vickers, D.F. Ransohoff, and G.S. Collins. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): Explanation and elaboration. *Https://Doi.Org/10.7326/M14-0698*, 162(1):1–73, 2015. doi:10.7326/M14-0698.

16 K.M. Nazi, C.L. Turvey, D.M. Klein, and T.P. Hogan. A decade of veteran voices: examining patient portal enhancements through the lens of user-centered design. *Journal of medical Internet research*, 20(7):e10413, 2018. doi:10.2196/10413.

17 N. RiahiNia, F. Shadanpour, K. Borna, and G.A. Montazer. Automatic keyword extraction using latent dirichlet allocation topic modeling: Similarity with golden standard and users' evaluation. *Human Information Interaction*, 9(3):1–22, 2022. doi:10.5555/944919.944937.

18 K. Skivington, L. Matthews, S.A. Simpson, P. Craig, J. Baird, J.M. Blazeby, K.A. Boyd, N. Craig, D.P. French, E. McIntosh, M. Petticrew, J. Rycroft-Malone, M. White, and L. Moore. A new framework for developing and evaluating complex interventions: update of medical research council guidance. *BMJ*, 374, 2021. doi:10.1136/BMJ.N2061.

19 T.A. Soriano, L.V. DeCherrie, and D.C. Thomas. Falls in the community-dwelling older adult: a review for primary-care providers. *Clinical interventions in aging*, 2(4):545–553, 2007. doi:10.2147/cia.s1080.

20 R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996. doi:10.1111/j.2517-6161.1996.tb02080.x.

21 VeiligheidNL. Infographic valongevallen 65-plussers 2022, 2022. URL: `https://www.veiligheid.nl/kennisaanbod/infographic/infographic-valongevallen-65-plussers-2022`.

22 H.M. Vu, L.H. Nguyen, H.L.T. Nguyen, G.T. Vu, C.T. Nguyen, T.N. Hoang, T.H. Tran, K.T.H. Pham, A. Latkin, Xuan Tran C., S.H.Ho B., C., and R.C.M. Ho. Individual and environmental factors associated with recurrent falls in elderly patients hospitalized after falls. *International Journal of Environmental Research and Public Health*, 17(7):2441, 2020. `doi:10.3390/ijerph17072441`.

23 L. Westerbeek, G.J. Bruijn, H.C. Weert, A. Abu-Hanna, S. Medlock, and J.C.M. Weert. General practitioners' needs and wishes for clinical decision support systems: A focus group study. *International Journal of Medical Informatics*, 168:104901, 2022. `doi:10.1016/J.IJMEDINF.2022.104901`.

24 L. Westerbeek, K.J. Ploegmakers, G.J. Bruijn, A.J. Linn, J.C.M. Weert, J.G. Daams, N. Velde, H.C. Weert, A. Abu-Hanna, and S. Medlock. Barriers and facilitators influencing medication-related cdss acceptance according to clinicians: A systematic review. *International Journal of Medical Informatics*, 152:104506, 2021. `doi:10.1016/J.IJMEDINF.2021.104506`.

25 C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021. `doi:10.1016/j.knosys.2021.106775`.

Real-Time Data-Driven Maintenance Logistics: A Public-Private Collaboration

Willem van Jaarsveld[1]
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Alp Akçay
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Laurens Bliek
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Paulo da Costa
Etsy Inc., Dublin, Ireland

Mathijs de Weerdt
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Rik Eshuis
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Stella Kapodistria
Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

Uzay Kaymak
Jheronimus Academy of Data Science, Eindhoven University of Technology, The Netherlands

Verus Pronk
Philips Research Laboratories, Eindhoven, The Netherlands

Geert-Jan van Houtum
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Peter Verleijsdonk
Department of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands

Sicco Verwer
Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, The Netherlands

Simon Voorberg
Department of Information Systems Supply Chain Management and Decision Analysis, NEOMA Business School, Rouen, France

Yingqian Zhang
Department of Industrial Engineering and Innovation Sciences, Eindhoven University of Technology, The Netherlands

Abstract

The project *"Real-time data-driven maintenance logistics"* was initiated with the purpose of bringing innovations in data-driven decision making to maintenance logistics, by bringing problem owners in the form of three innovative companies together with researchers at two leading knowledge institutions. This paper reviews innovations in three related areas: How the innovations were inspired by practice, how they materialized, and how the results impact practice.

2012 ACM Subject Classification General and reference → Surveys and overviews

Keywords and phrases Data, Maintenance, Logistics, Optimization, Research, Project

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.5

Funding This work was supported by the "Netherlands Organisation for Scientific Research" (NWO). Project: NWO Big data - Real Time ICT for Logistics. Number: 628.009.012.

Acknowledgements We want to thank all the involved private partners who have contributed to this project.

[1] Corresponding author

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 5; pp. 5:1–5:13
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Companies in maintenance logistics aspire to make better use of the increasing availability of real-time data from the many (inter)connected devices within the internet of things (IoT). Many such companies have in recent years taken a first key step in this direction by investing substantially in a data management infrastructure that ensures central and real-time availability of the raw data generated by the assets as well as key information on maintenance resources; e.g. the real-time location and status of field service engineers and spare parts, the status of repair centers, the availability of remote service engineers in the call center, etc. Companies are eager to leverage this investment to reduce cost and increase operational asset availability, by transitioning from traditional static maintenance logistics plans based on rigid task intervals to dynamic maintenance logistics policies fueled by real-time data.

Against this background, in 2018 the project *"Real-time data-driven maintenance logistics"* was started, as a collaboration between two leading knowledge institutions in the area of data-driven maintenance logistics (Eindhoven University of Technology and Delft University of Technology) and three companies at the forefront of innovations in maintenance logistics: Philips, Fokker Services (FS), and Dutch Railways (NS). The objective of this project according to its proposal was set in two steps:

1. The dynamic identification of actions from real-time data.
2. Organization of the dynamic execution of these actions by appropriately allocating the resources to accomplish them.

A different way of looking at it, based on the outcome of multiple consortium-wide discussions and following the design of an award-winning poster [14], would be the first step requiring a prediction that could be followed up by a prescription. The second step would design the process for this prescription. Hence, the outcome of our research would result in algorithms that identify the appropriate actions and a semi-structured data-driven process model that dynamically accomplishes such actions.

Big data played an important role, in the following sense. All three companies have large amounts of data available, and seek to make their operations data-driven. However, to achieve this, planning must become more nimble, requiring different planning algorithms and approaches. Hence, the general development of data-driven operations was the key motivation for the project.

To keep the project relevant, in the project execution we took a more flexible approach to identifying research topics. In particular, the project featured meetings with the entire consortium, approximately every 6 months. During those meetings, we reported the results of the previous period, as well as plans for the coming period. More importantly, companies provided feedback on those plans, and contributed ideas on how the various developments linked to challenges within each company. Those ideas enabled researchers to refine and make concrete the research directions set out in the project proposal. This led to various concrete research challenges, that were actively researched by various researchers funded by the project. Amongst others, we researched

1. How to include integer constraints in expensive optimization problems.
2. How to optimize information gathering when decision making.
3. How to learn from data to do maintenance prognostics and routing.

For each of these endeavors, we seek to highlight:

(i) How the challenge arose from the interaction with the participating companies.
(ii) How the challenge was formalized into research.

(iii) How data was used/collected, including adjustments and enrichments.

(iv) Which algorithms existed and which new algorithms were developed in the project.

(v) Experiments and results.

(vi) The learnings and practical validation with the companies.

We discuss these 6 elements for integer constraints in expensive optimization problems in Section 2, for optimized information gathering in Section 3, and for learning maintenance prognostics and routing in Section 4. We reflect and conclude in Section 5.

2 Integer constraints in expensive optimization problems

2.1 Domain

The Train Unit Shunting Problem (TUSP) is a complicated planning problem in railway operations, where trains are moved to a shunting yard to be maintained, cleaned and inspected, see Figure 1. The research group at NS has several algorithmic solutions available for this problem, each with its own set of parameters that need to be tuned correctly. However, tuning these parameters requires not just domain knowledge of railways, but also knowledge of the planning algorithms. Such broad knowledge is difficult to find in a team, let alone in one employee. Automated support for tuning the parameters of planning algorithms is therefore very relevant to achieve top performance in planning problems.

An issue is that tuning the parameters of planning algorithms, such as the algorithms used for TUSP, is very time consuming, and can be regarded as an expensive optimization problem due to the required computational resources. Predicting the performance of certain parameter values in advance would help in the tuning process. This can be done with surrogate-based optimization (SBO) techniques, which are particularly well suited for solving expensive optimization problems [17, Chapter 10]. These SBO techniques use machine learning to predict the performance of parameter values and to find the optimal values.

Traditional SBO techniques assume that parameters can be tuned to any real-valued number between a lower and upper bound. Some parameters in TUSP algorithms, however, are integer-valued, meaning that there is only a finite number of possibilities available for them. We call such a restriction an integer constraint. How to best deal with integer constraints in expensive optimization problems is an open research question.

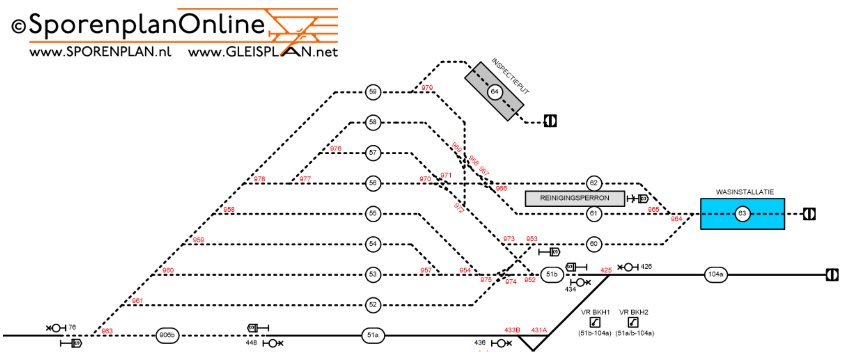


Figure 1 Example of a train shunting yard considered in the TUSP, from `www.sporenplan.nl`.

2.2 Modeling

Motivated by the problem of parameter tuning of TUSP algorithms, we study expensive optimization problems in general, particularly those with integer constraints. The objective (e.g., algorithm performance) is treated as a black-box and may suffer from noise, for example due to randomness in the TUSP algorithms or disturbances when executing the proposed solution. This way, the problem is reduced to tuning the parameters in such a way that the objective is maximized.

2.3 Data

We consider several simulators and parameter tuning problems that are expensive to run: a robust Traveling Salesman Problem, a wind farm simulator, a pipe shape simulator, an industrial gas filter simulator, a hospital simulator, and hyperparameter tuning for an XGBoost machine learning model. These have been implemented in an open-source software package, EXPOBench [5]. Since gathering data from these simulators is computationally expensive, all data resulting from the project has been made available open-source [7], providing a new big dataset to the community. Additionally, the parameters of several TUSP algorithms have been tuned automatically in MSc projects, namely two TUSP solvers [21, 20] and a TUSP instance generator [16].

2.4 Algorithms and experiments

We compared with the following baseline SBO techniques: Bayesian optimization with Gaussian processes [19], SMAC [15], HyperOpt [2], CoCaBO [18], and DONE [4]. The DONE algorithm has been adapted such that it can deal with integer constraints, leading to two new techniques that were developed during the project: IDONE [8] and MVRSM [6].

Experiments consisted of running different SBO techniques on different simulators and parameter tuning problems. An open-source software package (EXPOBench) was created for this purpose [5], and the resulting dataset was also publicly released.

2.5 Practice

Domain experts can now tune the parameters of their algorithms even when different types of parameters (continuous, integer) are involved, due to the newly developed techniques, and different solutions are available for domain experts to compare. Furthermore, we have obtained valuable insights into the relation between the type of SBO algorithm and the type of parameters in the problem, providing more guidelines on when to use which technique. The first results on algorithms from NS showed the potential for improvement in parameter tuning, but the algorithms need to be made more consistent before they can really benefit from the research. More research into the generalization aspects of the SBO algorithms is also required, to make sure they can operate in a variety of situations. Besides using the newly developed techniques at project partner NS in MSc projects, these techniques have also been applied in other places such as Redeia in Spain, and in other software packages such as fast CMA-ES [26]. The research has been presented at several companies, conferences and workshops.

3 Optimizing information gathering

3.1 Domain

Work in maintenance is often outsourced to specialized suppliers, and the process of finding and safeguarding the availability of such suppliers, which in turn enables those suppliers to engage in capacity management, is information- and knowledge-intensive [24, see also]. In particular, operators of high-tech equipment such as trains and aircraft send out quotations to a diverse range of potential suppliers. Processing such quotations is time consuming. For suppliers, not every quotation is sufficiently interesting to invest the time required to process it.

Most time is invested in gathering the information based on which the decision, to quote and what to quote, is made. There is no fixed procedure for this, since every quote is perceived to be unique. This information gathering costs time and money. Therefore, companies such as FS require an approach that helps them guide their employees during the quotation process, by either recommending when to stop since the quote indicates a likely loss for FS, or by recommending which information to keep collecting. Three key and interlocking challenges in the efficient processing of quotations are:

1. How to avoid the collection of information that is not useful.
2. How to ensure that sufficient information is gathered before the deadline for submitting a quote passes.
3. The order in which the information is collected should be according to what is the most valuable information to optimize the quote.

3.2 Modeling

The challenge is approached via a combination of conceptual modeling and optimization. From the conceptual modeling point of view, we adopt an industrial modeling language for dynamic information-centric business processes denoted by CMMN (Case Management Modeling and Notation). Choosing an industrial standard for our approach helps transfer the research into practice. We introduce an approach that can model decision-intensive processes, or more specifically quotation processes, in CMMN. Moreover, we introduce additional inputs such as an information structure which represents the business case that is being developed as part of submitting a quote, i.e. the projected profit as a function of the information retrieved. The result of this part of the model is defined as an Optimizable Decision-Intensive Process (ODIP) consisting of a CMMN representation of the process according to a certain set of constraints that allow for the following approach.

To optimize information retrieval in this ODIP, we develop an approach to convert the CMMN model into a Markov Decision Process (MDP). An MDP is specifically useful in sequential decision processes with uncertain environments, in which the effect of decisions is also uncertain. Given the uncertain outcome of a quotation and the uncertain input of information, an MDP is highly appropriate to model such quotation process optimization problems and more in general ODIPs. Together, the ODIP model and the MDP derived from it, result in a decision support policy that supports the knowledge worker in making the best decision while going through a quotation process. This decision support policy has also resulted in an online demonstration tool in which we show the effectiveness of the complete approach.

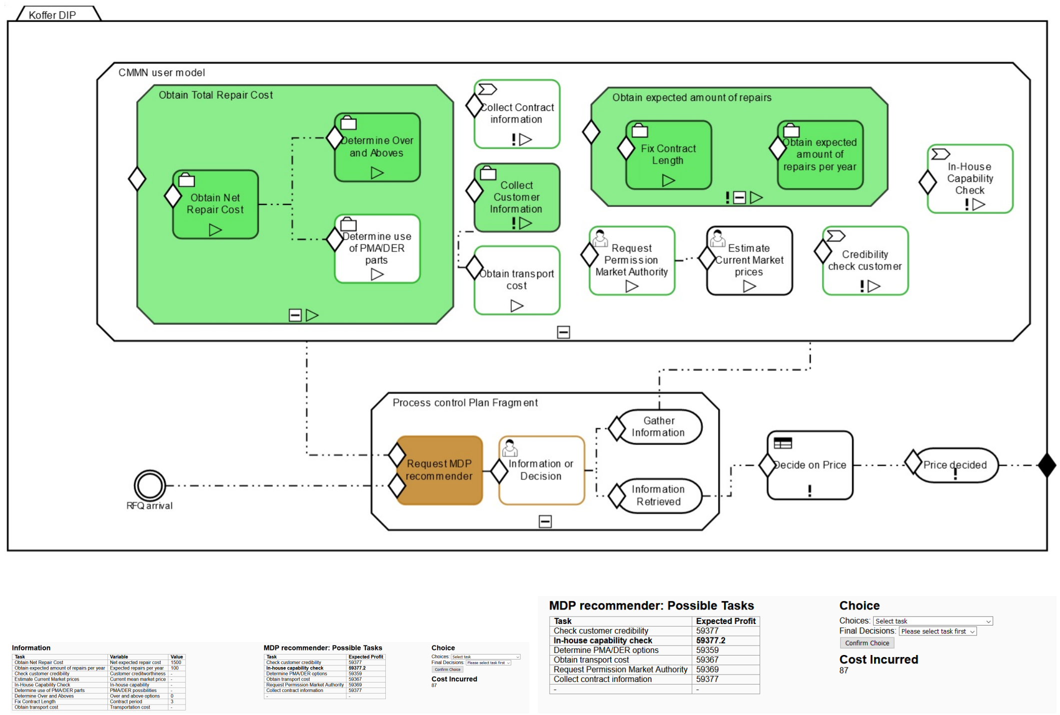


Figure 2 A visualization of the demonstration tool that provides decision support to knowledge workers in a quotation process. (`OptimizingInformationGathering`).

3.3 Algorithms

To optimize the information gathering, we adopt an algorithm for solving the MDP generated from the CMMN process. Using the output of the algorithm, users are recommended to stop the process (since a loss is expected or more information will not result in higher profits) or to continue and collect a specific piece of information. The support is highly flexible: Whenever the user overrides a recommendation, for example because of tacit knowledge based on experience that is not clear from the data, the algorithm adapts and yields new recommendations appropriate for the path adopted by the user. For small size ODIPs, we are able to find the optimal solution since there is a finite horizon to these problems, using backward recursion. For large size problems, we have introduced a deep reinforcement algorithm to still find feasible and good-performing policies.

3.4 Data and experiments

The main inspiration that led us to this model comes from the quotation optimization problem at FS. Based on their definition of the process, their information inputs (which were simplified to keep the true information anonymous and the solution tractable) and their goals, we have constructed a business case that served as a first experiment (see Figure 2). In this figure, one sees a decision process where multiple pieces of information have already been collected (filled green) and where the green outlined cases are information that can be collected if deemed necessary. The bottom right part of the tool gives an insight in the expected profit when collecting a certain type of information. This experiment allowed us to show the feasibility of the full approach for an existing decision-intensive process. Subsequently, we show the effectiveness and efficiency of our policies in a full-factorial experiment vis-a-vis relevant benchmarks.

3.5 Practice

Using the developed approach, considerable time can be saved when developing quotations. The challenging part is the development of concrete business cases; ideas for this have been developed both at FS and at companies not participating in the project, such as a large logistics service provider in the southeast of the Netherlands. Another issue is scalability. In solving the optimization problem as a MDP, a challenge is that the number of states grows exponentially in the size of the problem. In successive research, we are actively exploring various strategies to tackle this. In one direction, we decompose an ODIP into subproblems based on decision hierarchy. Large decision problems often contain multiple subdecisions that allow us to use this approach. In the other direction, we introduce solution methods that can deal with significantly larger state spaces such as deep reinforcement learning. Such methods increase the applicability of the model to a wider range of problems.

4 Learning maintenance prognostics and routing

4.1 Domain

Many prediction and optimization problems arise to obtain policies that maintain a network of industrial assets with minimum maintenance and travel costs. Traditionally, the prediction and optimization problems are decoupled from each other, and prediction uncertainties are incorporated into the decision making models; see [1] for a recent example in the context of maintenance. Furthermore, the methods for solving the prediction and optimization problems rely on the modeling and solution capabilities of decision makers, which can be inaccurate or inefficient for the problems at hand. The motivation of this work is to move from this traditional view to the adoption of data-driven methods based on machine learning leading to accurate and efficient solutions. The scope of our work includes the problems in two distinct areas, prognostics and routing, that come together in the maintenance of a network of industrial assets.

4.1.1 Prognostics

The field of prognostics focuses on predicting the remaining useful life (RUL) of equipment, which is critical for optimizing maintenance schedules and minimizing downtime. One of the primary challenges in prognostics is the difficulty in determining the best time to repair a given asset. In many cases, sensor data from equipment is available; however, parsing this information can be time-consuming and complex due to the high volume and variety of data streams. Furthermore, for newly introduced equipment, run-to-failure data is often unavailable, making it even more challenging to accurately predict the RUL. On the other hand, historical (run-to-failure) data may exist for older assets, providing valuable insights for estimating their RUL. Hence, the development of advanced prognostic algorithms capable of leveraging both sensor data and historical data is crucial for optimizing repair schedules and improving the overall reliability of equipment.

4.1.2 Routing

In the context of equipment maintenance and repair, routing can be formulated as a Traveling Salesman Problem (TSP) where the goal is to determine the most efficient route for visiting a network of assets that require inspection or maintenance. Figure 3 illustrates the solution of a TSP instance. Given the combinatorial nature of the TSP, finding an optimal solution is computationally intractable for large-scale problems. Therefore, we propose adopting a machine learning (ML) perspective to learn improvement heuristics, which can be used to search near-optimal tours. By employing ML techniques such as reinforcement learning, we aim to develop algorithms capable of efficiently exploring the possible solutions and converging to high-quality solutions for the routing problem. This approach not only enables more effective utilization of available resources but also the incorporation of asset prognostics into the routing decision making process.

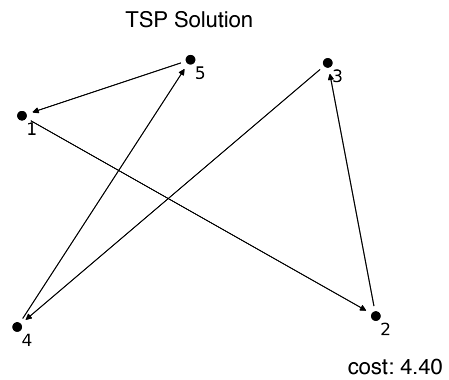


Figure 3 A Traveling Salesman Problem solution. Indices represent the order in which nodes, representing locations in the network, are visited.

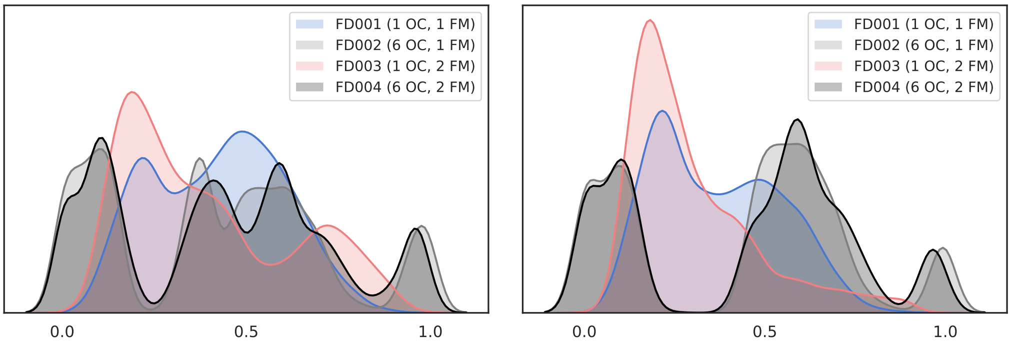
4.2 Modeling

In the field of prognostics, our focus is on developing methods that require minimal human intervention and do not necessitate prior knowledge of the underlying failure mechanisms. The proposed methods are designed to leverage historical run-to-failure data, which is assumed to be available for older equipment, and sensor data for the assets of interest. By modeling this setup as a transfer learning problem, we aim to use the assets with both sensor and run-to-failure data to improve prognostics assets with only sensor data [13]. To effectively capture dependencies in the sensor data, we employ machine learning techniques [12].

In the context of routing, we initially examine classic TSP improvement heuristics as a starting point and baseline for our proposed methods. We hypothesize that a look-ahead policy based on simple operators, can outperform a short-sighted one. Our objective is to learn such policies autonomously, relying solely on executing actions and observing their long-term results. Moreover, we do not have access to an exact solver, as generating optimal tours can become computationally intractable for large-scale problems [9, 10]

4.3 Data and experiments

In the context of prognostics, we utilize simulated data from aircraft turbofan engines operating under various conditions and fault modes. The experimental setup consists of four datasets (i.e., FD001-FD004), each containing different combinations of fault modes and operating conditions. Figure 4 illustrates two normalized sensor values just before a failure occurs. In each experiment, one dataset is designated as the source domain, and we attempt to learn the RUL of the remaining three datasets (target domains), which have different fault modes and operating conditions. This approach allows us to test the model's effectiveness in transferring knowledge across operating conditions and fault modes. For each target domain dataset, we assume no access to the observed RUL of the assets and attempt to make predictions based solely on the sensor data from the target domain and observed RUL information from the source domain.



(a) Sensor A. **(b)** Sensor B.

Figure 4 Distribution of normalized sensor values before a failure. Sensor distributions differ for the same assets under different operating conditions and fault modes.

For the routing experiments, we consider simulated data containing TSP instances of various sizes (20, 50, 100). We construct a simulator capable of handling a batch of TSP instances and making decisions corresponding to improvement operators based on 2-edge swaps. Data is collected following policies in the simulator, whereby a sequence of operators creates a history of TSP solutions and selected operators for a batch of TSP instances. We train and evaluate the proposed methods for different TSP sizes, learning policies for a 50-node TSP and evaluating on a 100-node TSP to assess how well the model generalizes to different sizes. We repeat a similar setup for other routing problems such as the Vehicle Routing Problem. Additionally, we test the proposed method on real-world TSPLib instances, which are not seen during the training process, in order to evaluate the model's generalization to instances with different location distributions.

4.4 Algorithms

In the field of prognostics, we develop a deep transfer learning algorithm which means we apply deep learning techniques in transfer learning algorithms. To evaluate the performance of the proposed algorithm, we compare it with various methods, such as transfer component analysis (TCA), and correlation analysis (CoRaL). Furthermore, we investigate the effectiveness of the proposed algorithm in comparison to methods trained only on the source domain (Source-only), and standardized data approaches.

In the routing context, we develop a deep reinforcement learning algorithm designed to select 2-edge swaps for the TSP, the multiple Traveling Salesman Problem, and the Vehicle Routing Problem. A 2-edge swap means that we swap two parts of the initial route and check if this improves the performance. We compare the performance of this algorithm against several benchmarks, including exact solvers, classic heuristics, and both supervised and reinforcement learning methods. This extensive comparison allows us to assess the relative strengths and weaknesses of our proposed approach, as well as its potential for application in real-world routing and scheduling problems.

4.5 Practice

In the context of prognostics, domain experts can benefit from our proposed deep transfer learning algorithm when faced with similar business cases, such as the presence of RUL labeled data for older assets and the need to obtain predictions for new assets that lack observed failures. As part of the project, [25] has applied our method in a use case at FS, and investigated how it can improve health state predictions of a specific aircraft component under varying working conditions. The strength of our method lies in its ability to be applied even in the absence of information about the underlying degradation of assets. Through the evaluation of test data, we have demonstrated that our proposed method outperforms other off-the-shelf transfer learning algorithms and yields more accurate predictions than training solely on the source domain.

In the realm of routing problems, our proposed heuristics offer domain experts an effective solution for TSP-like problems. The learned policies outperform other hand-crafted heuristics while operating under a similar computational budget. A notable advantage of our algorithm is its capacity to be applied in batches, enabling the efficient simultaneous solving of multiple instances. Experimental results further reveal that our algorithm can be employed to solve more general TSP instances, and can be adapted to fit specific instance distributions depending on the application, such as drilling problems, cargo delivery, maintenance scheduling and chip design.

Motivated by the success of deep reinforcement learning for solving the TSP efficiently, the approach has been extended for a real-life service-logistics use case of Philips, where the routing of a field service engineer is optimized in the presence of imperfect alerts on the health condition of physical assets [11]. Based on the discussions with Philips, we developed the Dynamic Traveling Maintainer Problem with Alerts (DTMPA) on asset networks. In the DTMPA, we rank the quality of information retrieved from the alert using various information levels. We propose a wide range of heuristics for the DTMPA to cover each information level. The numerical results show that the deep reinforcement learning heuristic outperforms the others, requiring least information. The DTMPA is extended to the case of multiple field service engineers who must coordinate actions to minimize asset unavailability whilst maximizing coverage to anticipate future events in the network [22]; see Figure 5 for a schematic depiction of that case. We employ an iterative deep reinforcement learning algorithm to directly improve sophisticated dispatching strategies by learning e.g. a repositioning strategy from simulated data. A key advantage of this approach is scalability since data collection can be distributed over multiple compute nodes, which enables us to solve industrial-scale problems in a reasonable time. Numerical experiments based on the Dutch hospital network show that the algorithm quickly produces near-optimal policies and the trained policies demonstrate robustness against minor model modifications.

5 Conclusion and reflection

In this paper, we discussed the various research endeavors that were part of the project *"Real-time data-driven maintenance logistics"*. Reflecting on the process, we next list a few observations and learnings we obtained from this project.

(a) Failure: When assets fail, available engineers must be dispatched efficiently.

(b) Alert: When an alert is issued, a risk urgency assessment decides whether to dispatch an engineer.

(c) Repositioning: Idle engineers are proactively repositioned to be closer to future alerts and failures.

Figure 5 A multi-maintainer service-logistics model: Service engineers respond to IoT-generated alerts of various degrees of severity that must be appropriately ranked considering risk, urgency and opportunity.

First of all, working together with companies worked well for this type of project. The companies each bought into the project by promising a monetary contribution. As a consequence, they each had the objective to get something out of the project, and project participants from the companies were aligned on this with senior management. As a consequence, company representatives were interested in the results and actively participated in project meetings. A good example of this is the award-winning poster [14], which was built based on intensive consortium-wide discussions.

The company input steered the direction of the research on various occasions, and in some cases, this led to research that is challenging theoretically, but that also has a high practical value. A key example is the work on transfer learning [13] for predicting remaining useful life via deep domain adaptation. This work was inspired by a comment from one of the companies that they had ample data for older systems, but not so much for new ones, and that it would be useful to be able to use data for the old systems to predict failures for the new systems. This work unlocked a whole stream of literature on precisely that subject, exactly because it is theoretically challenging. Also, these ideas turned out to be applicable in the company setting [25]. Similarly, the use case on efficiently dealing with requests for quotations was a basis for work that is interesting both in practice and in theory [24, 23].

Some work was useful in ways that were not initially foreseen. For example, the surrogate models discussed in Section 2 were not designed with problems in maintenance logistics in mind, but turned out to be really useful in that context. Relatedly, we note that while each of the junior researchers on the project had their own focus, a substantial overlap both in methodological interests and in problem domain enabled them to learn from each other, which strengthened the entire team and enabled us to deliver interesting interdisciplinary research insights, e.g. the comparison of various traditional optimization techniques and deep reinforcement learning by [10], and the first Traveling Salesman Problem competition [3] which similarly bridges disciplines and was organized via a cooperation within the consortium.

If we look back at the two-step objective of this project we can conclude that we have achieved our goals in both steps. In Sections 2 and 4, we have shown some of the achievements related to the first step regarding identifying actions from real-time data. Also, in Sections 3 and 4, we show how actions can be implemented into a decision making process where we aim at dynamically executing them.

6 Guidance for future collaborative projects

Reflecting on our project's journey, several key strategies stand out as instrumental for the success of similar collaborative ventures. A critical element was ensuring company buy-in, particularly through promised contributions. This commitment significantly boosted their ongoing engagement and helped align their objectives with the project's goals. Regular, well-structured consortium meetings also played a vital role. These gatherings maintained company interest and facilitated real-time alignment with their evolving needs, proving essential for sustained collaboration.

Additionally, our project's flexible planning approach, which emphasized broader goals over specific objectives, allowed us to nimbly adapt to changes and unexpected developments within the companies. This flexibility was key in marrying theoretical exploration with practical applicability, ensuring the project's relevance and responsiveness.

These insights offer valuable lessons for future research initiatives aiming to bridge the gap between academia and industry. Strategic engagement, consistent communication, and adaptable planning emerge as fundamental components for fostering effective partnerships and achieving meaningful outcomes in such collaborative settings.

References

1 Alp Akcay. An alert-assisted inspection policy for a production process with imperfect condition signals. *European Journal of Operational Research*, 298(2):510–525, 2022. doi:10.1016/J.EJOR.2021.05.051.

2 James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 115–123. PMLR, 17–19 June 2013. URL: http://proceedings.mlr.press/v28/bergstra13.html.

3 L Bliek, P da Costa, R Refaei Afshar, Y Zhang, T Catshoek, D Vos, S Verwer, and F Schmitt-Ulms. The first AI4TSP competition. https://paulorocosta.gitbook.io/ai4tsp-competition/.

4 L. Bliek, H. R. G. W. Verstraete, M. Verhaegen, and S. Wahls. Online optimization with costly and noisy measurements using random Fourier expansions. *IEEE Transactions on Neural Networks and Learning Systems*, 29(1):167–182, 2018. doi:10.1109/TNNLS.2016.2615134.

5 Laurens Bliek, Arthur Guijt, Rickard Karlsson, Sicco Verwer, and Mathijs De Weerdt. EXPObench: Benchmarking surrogate-based optimisation algorithms on expensive black-box functions. *arXiv preprint arXiv:2106.04618*, 2021. arXiv:2106.04618.

6 Laurens Bliek, Arthur Guijt, Sicco Verwer, and Mathijs De Weerdt. Black-box mixed-variable optimisation using a surrogate model that satisfies integer constraints. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '21, pages 1851–1859, New York, NY, USA, 2021. Association for Computing Machinery. doi:10.1145/3449726.3463136.

7 Laurens Bliek, Arthur Guijt, Sicco Verwer, Mathijs De Weerdt, and Rickard Karlsson. Raw data of the EXPensive Optimization benchmark library (EXPObench). Available at https://doi.org/10.4121/14247179.v2, 2021.

8 Laurens Bliek, Sicco Verwer, and Mathijs De Weerdt. Black-box combinatorial optimization using models with integer-valued minima. *Annals of Mathematics and Artificial Intelligence*, 89:639–653, 2021. doi:10.1007/s10472-020-09712-4.

9 Paulo da Costa, Jason Rhuggenaath, Yingqian Zhang, and Alp Akcay. Learning 2-opt heuristics for the traveling salesman problem via deep reinforcement learning. In *Asian Conference on Machine Learning*, pages 465–480. PMLR, 2020.

10 Paulo da Costa, Jason Rhuggenaath, Yingqian Zhang, Alp Akcay, and Uzay Kaymak. Learning 2-opt heuristics for routing problems via deep reinforcement learning. *SN Computer Science*, 2:1–16, 2021.

11 Paulo da Costa, Peter Verleijsdonk, Simon Voorberg, Alp Akcay, Stella Kapodistria, Willem Van Jaarsveld, and Yingqian Zhang. Policies for the dynamic traveling maintainer problem with alerts. *European Journal of Operational Research*, 305(3):1141–1152, 2023. `doi:10.1016/J.EJOR.2022.06.044`.

12 Paulo Roberto De Oliveira da Costa, Alp Akcay, Yingqian Zhang, and Uzay Kaymak. Attention and long short-term memory network for remaining useful lifetime predictions of turbofan engine degradation. *International Journal of Prognostics and Health Management*, 10(4), 2019.

13 Paulo Roberto de Oliveira da Costa, Alp Akçay, Yingqian Zhang, and Uzay Kaymak. Remaining useful lifetime prediction via deep domain adaptation. *Reliability Engineering & System Safety*, 195:106682, 2020. `doi:10.1016/J.RESS.2019.106682`.

14 Paulo De Oliveira Da Costa et al. Real-time data-driven maintenance logistics. 3rd Place Commit2Data poster competition at ICT.OPEN, March 2019.

15 Frank Hutter, Holger H Hoos, and Kevin Leyton-Brown. Sequential model-based optimization for general algorithm configuration. In *International conference on learning and intelligent optimization*, pages 507–523. Springer, 2011. `doi:10.1007/978-3-642-25566-3_40`.

16 R. Martens. Controlling TUSP features during instance generation via Bayesian optimization. Eindhoven University of Technology, Master's thesis, 2022.

17 Joaquim R. R. A. Martins and Andrew Ning. *Engineering Design Optimization*. Cambridge University Press, 2021. `doi:10.1017/9781108980647`.

18 Binxin Ru, Ahsan Alvi, Vu Nguyen, Michael A. Osborne, and Stephen Roberts. Bayesian optimisation over multiple continuous and categorical inputs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8276–8285. PMLR, 13–18 July 2020. URL: `http://proceedings.mlr.press/v119/ru20a.html`.

19 B. Shahriari, Kevin Swersky, Ziyu Wang, R. Adams, and N. D. Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. `doi:10.1109/JPROC.2015.2494218`.

20 L. J. Van den Nieuwelaar. Automatic algorithm configuration with search heuristics for the train unit shunting problem. Eindhoven University of Technology, Master's thesis, 2021.

21 L. Van der Knaap. Contextual hyperparameter optimization for the train unit shunting problem. Delft University of Technology, Master's thesis, 2021.

22 Peter Verleijsdonk, Willem van Jaarsveld, and Stella Kapodistria. Scalable policies for the dynamic traveling multi-maintainer problem with alerts. *Eur. J. Oper. Res.*, 319:121–134, 2024. `doi:10.1016/J.EJOR.2024.05.049`.

23 S Voorberg, W Van Jaarsveld, R Eshuis, and GJ Van Houtum. Information acquisition for service contract quotations made by repair shops. *European Journal of Operational Research*, 305(3):1166–1177, 2023. `doi:10.1016/J.EJOR.2022.06.048`.

24 Simon Voorberg, Rik Eshuis, Willem Van Jaarsveld, and GJ Van Houtum. Decisions for information or information for decisions? Optimizing information gathering in decision-intensive processes. *Decision Support Systems*, 151:113632, 2021. `doi:10.1016/J.DSS.2021.113632`.

25 H.G.P. Wismans. Domain adaptation for prognostics in the aerospace industry. Eindhoven University of Technology, Master's thesis, 2019.

26 Dietmar Wolz. FCMAES - a Python-3 derivative-free optimization library. Available at `https://github.com/dietmarwo/fast-cma-es`, 2022.

WheelPower: Wheelchair Sports and Data Science Push It to the Limit

Riemer J. K. Vegter[1]
University Medical Center Groningen,
Center for Human Movement Sciences,
University of Groningen, The Netherlands

Rowie J. F. Janssen
University Medical Center Groningen,
Center for Human Movement Sciences,
University of Groningen, The Netherlands

Marit P. van Dijk
Department of Biomechanical Engineering,
Delft University of Technology, The Netherlands

Marco J. M. Hoozemans
Department of Human Movement Sciences,
Faculty of Behavioural and Movement Sciences,
Vrije Universiteit Amsterdam, The Netherlands

Dirkjan H. E. J. Veeger
Department of Biomechanical Engineering,
Delft University of Technology, The Netherlands

Han J. H. P. Houdijk
University Medical Center Groningen,
Center for Human Movement Sciences,
University of Groningen, The Netherlands

Luc H. V. van der Woude
University Medical Center Groningen,
Center for Human Movement Sciences,
University of Groningen, The Netherlands

Monique A. M. Berger
Faculty of Health Nutrition and Sports,
The Hague University of Applied Sciences,
The Netherlands

Rienk M. A. van der Slikke
Faculty of Health Nutrition and Sports,
The Hague University of Applied Sciences,
The Netherlands

Sonja de Groot
Department of Human Movement Sciences,
Faculty of Behavioural and Movement Sciences,
Vrije Universiteit Amsterdam, The Netherlands

Abstract

Paralympic wheelchair athletes solely depend on the power of their upper-body for their on- court wheeled mobility as well as for performing sport-specific actions in ball sports, like a basketball shot or a tennis serve. The objective of WheelPower is to improve the power output of athletes in their sport-specific wheelchair to perform better in competition. To achieve this objective the current project systematically combines the three Dutch measurement innovations (WMPM, Esseda wheelchair ergometer, PitchPerfect system) to monitor a large population of athletes from different wheelchair sports resulting in optimal power production by wheelchair athletes during competition. The data will be directly implemented in feedback tools accessible to athletes, trainers and coaches which gives them the unique opportunity to adapt their training and wheelchair settings for optimal performance. Hence, the current consortium facilitates mass and focus by uniting scientists and all major Paralympic wheelchair sports to monitor the power output of many wheelchair athletes under field and lab conditions, which will be assisted by the best data science approach to this challenge.

2012 ACM Subject Classification Social and professional topics → People with disabilities

Keywords and phrases Paralympic sports, Wheelchair sports, Power

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.6

1 Project overview

Paralympic wheelchair athletes solely depend on the power of their upper-body for their on-court wheeled mobility as well as for performing sport-specific actions in ball sports, like a basketball shot or a tennis serve. Wheeled mobility is dependent on power output

[1] Corresponding author

and can be measured in a standardized lab environment and to a certain extend during training and competition. Wheelchair athletes adhere to very intense training programs, but they until recently lacked sport-specific tools for objective monitoring of power production, essential to performance [5]. Moreover, different wheelchair sports are performed by a relatively small group of athletes with a plethora of different impairments due to trauma or disease (Paralympic.org). This has resulted in limited and scattered scientific knowledge directly applicable to their sports practice [1]. For instance, only recently the extra peak power necessary to propel the wheelchair, while simultaneously holding a tennis racket, was shown [3]. Hence, little is known about the biomechanical and physiological demands of the wheelchair sports, impeding the possibility to provide coach and athlete with essential feedback to achieve both individualized and team performance optimization through training, wheelchair design and fitting, and game strategy.

Fortunately, The Netherlands already has a strong research history on the use of manual wheelchairs by multiple knowledge partners, like VU Amsyterdam, TU Delft, University op Groningen and The Hague University of Applied Sciences. Together they approached the Paralympic Wheelchair Sports, represented by NOC*NSF and their individual sports federations. Together we opened the discussion and formulated the correct research questions, endorsed by all parties. Multiple sessions were organised with all stakeholders invited. Bringing together the needs of all partners and understanding each other has been one of the challenges already before, but also during the project. For example, the needs of coaches and players must be balanced with time that was asked for participating in structural measurements over the season, leading up to big tournaments. Therefore, seeing the immediate benefits and feeling that it improves their preparation was and is critical to the project.

More discussions and improved mutual understanding have already been benefits of the WheelPower project and have helped to make a stronger connection between Science and wheelchair sports. Given the strength of having multiple Paralympic wheelchair sports (wheelchair-basketball, -tennis, -rugby, -racing and paratriathlon) buy-in on this project, it became also more interesting to our commercial partners, who saw opportunities to improve their commercial products based on the expected generated knowledge. Similarly, the rehabilitation centers perceived the projects" possibility for knowledge translation to patient rehabilitation as very useful and became consortium partner. In total 17 partners signed the consortium agreement, which was a major effort, but has been instrumental to the strength of the project.

2 Problem addressed

The objective of WheelPower is to improve the power output of athletes in their sport-specific wheelchair to perform better in competition. To achieve this objective the current project systematically combines the three Dutch measurement innovations (WMPM, Esseda wheelchair ergometer, PitchPerfect system) to monitor a large population of athletes from different wheelchair sports resulting in optimal power production by wheelchair athletes during competition. The data will be directly implemented in feedback tools accessible to athletes, trainers and coaches which gives them the unique opportunity to adapt their training and wheelchair settings for optimal performance.

The main research question is: what are the performance indicators for optimal power production by wheelchair athletes during competition and how are they improved using continuous performance monitoring, combined with direct feedback? The sub-questions are:

1. How do the power demands, monitored during training and competition, compare to the maximum measured power production capabilities of wheelchair athletes?
2. What is the biomechanically optimal technique for efficient power transfer from the upper-body into the wheelchair-sport specific goals? For instance, how does timing of subsequent rotation of (upper-)body segments affect the efficiency of power transfer when pushing the hand rims, or when hitting the ball with the racket during wheelchair tennis?
3. How does direct feedback on the performance indicators for power production in wheelchair sports actions improve the competition performance of wheelchair athletes?

Multiple reasons explain the barriers that have caused wheelchair sports to fall behind in performance monitoring relative to other (wheeled) sports. First, upper-body cyclic physical activity is fundamentally different from lower-body tasks such as cycling, hampering knowledge translation from able-bodied sports [9]. Unlike the legs, the arms are essentially equipped for apprehension and 3D-manipulation of objects rather than heavy work such as load bearing. Therefore, the arms need constant joint stabilizing muscle activation, while the diversity of small muscle units has a low total muscle mass and thus power output, making them more injury prone [16]. Second, the mode of hand rim wheelchair propulsion is categorically different from any other form of ambulation, since during each propulsion cycle both hands need to bimanually couple to rotating rims outside the visual field in a discontinuous propulsion mode [17]. Third, the wheelchair design and sport-specific nature make the use of regular power sensors (e.g., Powertap [18], SRM [2]) as used in cycling, unfeasible. In addition, force instrumented hand-rims (e.g., SmartWheel [10], Optipush [6]) as used in more fundamental wheelchair research, are bulky and have too much impact on the mass and inertia of the light-weight sports wheelchair. Last, as stated earlier, the heterogeneity of wheelchair athletes – among others due to variation in impairments – has previously limited quantitative large sample size studies that generalize research insights to the whole group of wheelchair athletes.

The above-mentioned gaps of 1) monitoring power output in the lab and on the field in a 2) large sample of wheelchair athletes are addressed in WheelPower as follows. From different perspectives the Dutch research groups have shown leadership in the development of new technologies applicable to wheelchair sports to obtain objective performance measures, such as external power output in the field and in the lab (Figure 1). First, the Wheelchair Mobility Performance Monitor (WMPM) is an Inertial Measurement Unit (IMU)-based field solution to measure the overall on-court power demands during training and competition [14]. Second, the Esseda wheelchair ergometer is a high-resolution measurement tool to analyse how power is optimally delivered by an athlete under standardized repeatable conditions in their personal sports wheelchair [4]. Third, the PitchPerfect system is also an IMU-based tool, originally developed for pitching in baseball but also employable in wheelchair sports, to provide coaches and athletes of feedback on motion timing and power production through the kinetic chain in throwing and hitting [11]. In this project, the methodology of the PitchPerfect will provide insight in upper extremity movements and joint angles during propulsion and other wheelchair sport related upper-extremity activities. By implementing these three new technologies in all five Paralympic wheelchair sports, power output will be monitored in a large group of wheelchair athletes.

In conclusion, the current consortium facilitates mass and focus by uniting scientists and all major Paralympic wheelchair sports to monitor the power output of many wheelchair athletes under field and lab conditions, which will be assisted by the best data science approach to this challenge.

Figure 1 Figure 1: The Wheelchair Mobility Performance Monitor (WMPM) is an Inertial Measurement Unit (IMU)-based field solution to measure the overall on-court power demands during training and competition. The Esseda wheelchair (WC) ergometer is a high- resolution measurement tool to analyse how the power is optimally delivered by an athlete under standardized repeatable conditions in their personal sports. Third, the PitchPerfect system is also an IMU-based tool, originally developed for pitching in baseball but also employable in wheelchair sports, especially wheelchair tennis, to provide coaches and athletes feedback on motion timing and power production through the kinetic chain in throwing and hitting.

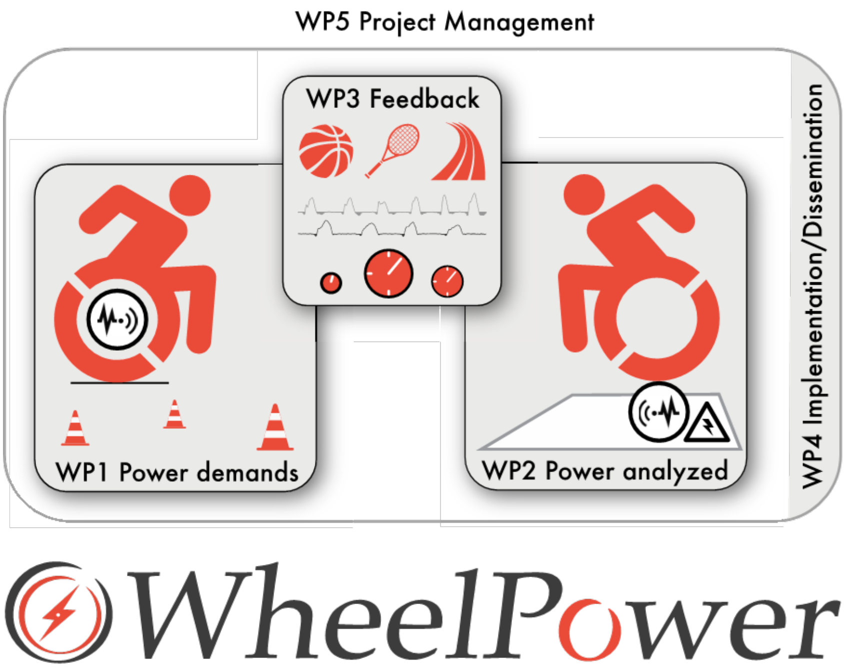


Figure 2 Figure 2: The WheelPower work packages (WP) as part of the consortium design.

3 Project consortium

Three main work packages form the core of the project (Figure 2). One Post-doc-and two PhD researchers are the main persons executing the research, i.e., perform the measurements and write scientific paper as well as provide feedback of the outcomes and discuss these with coaches, players and embedded scientist. They all function within the specific consortium teams with assigned work package leaders. To coordinate the whole project a three weekly meeting structure has been put in place with an alternating big and small team assembly. In the small team the project-leader talks with the three main researchers and their daily supervisors, while the other meeting assembles everybody involved from the research partners.

To maintain good communication with all people involved, a yearly meet-up is organised, with additional discussion with the strategic partners from the different Paralympic sports. For a wider audience a regular newsletter is maintained coupled to our website (`www.wheelpower.online`). Social media through Instagram, Twitter and LinkedIn have been very active and well followed. Especially athletes sharing their measurements through Instagram has been a strong addition to the visibility of the project and has helped to motivate other athletes to participate. Twitter and LinkedIn have been very useful for international visibility and making new connections to interested parties. The partners formalised in the WheelPower consortium have also extended and intensified their formal collaboration around Paralympic Science Support, not only for wheelchair sports, but now for alle Paralympic disciplines. This consortium now has a dedicated coordinator positioned at Kenniscentrum Sport en Bewegen and will help strengthen what was gained through the WheelPower project.

Regarding knowledge sharing and implementation we adhere to Open Science principles. All scientific papers are open access available, as well as most of the data analyses tools (`https://pypi.org/project/worklab/`).

4 Key results

The main outputs of WheelPower are (1) the addition of the power output to the WMPM (used in training and competition), (2) the development of test protocols to measure the (an)aerobic capacity of elite wheelchair athletes and (3) the applications of the WMPM during competition.

The first work packages add the estimation of power output to the already existing WMPM (Figure 3). This project started in a lab environment on an extra-large treadmill and several measurement devices were combined to objectively measure the power output produced by the individual. Based on these "golden standard" measurements, three IMU sensors (on the wheel, frame and trunk) were trained with machine learning and provided an accurate estimate of the power output. With the addition of power output to the WMPM, it becomes possible to monitor the training load of the players throughout the season to prevent injuries and perform maximally during major competitions [15].

The second work package ensures the successful implementation of wheelchair-specific anaerobic and aerobic exercise testing in all five Paralympic wheelchair sports. The standardized, yet individualized test protocols to test these maximal (an)aerobic capacities have been developed in WheelPower (Figure 4). First, an extensive literature review was written that provided an overview of all studies that addressed the wheelchair-specific anaerobic and/or aerobic exercise capacity [7]. After that, these different test protocols were synthesized into a test protocol that can assess the exercise capacity in a diverse group of wheelchair athletes, using a standardized, yet individualized way [8]. After validating these test protocols, they have been used for regular monitoring of Dutch athletes from the five Paralympic wheelchair

Figure 3 Figure 3: Wheelchair racing measurements on a treadmill for sensor-based power estimations.

sport disciplines. For example, the elite wheelchair tennis players monitor their (an)aerobic performance bi-annually and in cooperation with trainers, embedded scientists and coaches, training of the athlete is optimised. Another example can be found in the wheelchair set-up: a wheelchair rugby athlete improved his seating and asymmetry with a different seating position and a wheelchair triathlete is now wheeling with a different size of hand rim. The implementation of these test protocols worked because WheelPower closely collaborated with coaches and embedded scientist from the national teams.

The third work package implemented the WMPM for daily monitoring during important competitions (Figure 5 and 6) [13, 12]. To improve the evidence-based classification system of wheelchair rugby, measurements have been conducted at the World Championships of wheelchair rugby. Several athletes from different national teams performed standardized field tests and were monitored during the actual competition. By combining these two, we aim to get more insight in the relation between coordination impairments and performance. For instance, the role of muscle strength on wheelchair tennis performance is investigated to improve athlete classification, in collaboration with the International Tennis Federation. Lastly, in cooperation with Basketball Experience NL, a feedback system is developed which makes it possible to present live performance feedback from the IMU sensors directly to the audience or as a live-stream overlay during a match.

5 Further steps

Multiple future steps for the results and products of the WheelPower project are foreseen.

The first is making sure that the products and knowledge of the WheelPower project have a sustainable future in the sports they were developed for. To that end, Sports Data Valley is an important automation partner that will help secure the infrastructure and analyses tools. That way, in the future every athlete, coach, sports-professional and research will have a privacy-proof platform where the performance portfolio is stored, maintained and accessed to keep profiting from participation.

Secondly, the addition of more athletes from the same sports and other wheelchair sports (like wheelchair handball), are foreseen, to help them profit themselves from the analyses, but also to extend the database for research. More questions about for instance talent development and longitudinal performance and injury pathways still need to be answered.

Additionally, the step to rehabilitation and activities of daily living seems straightforward. Like Formula-1, the Paralympic athletes are prime examples of the physical and mental potential of manual, wheelchair users. Translation of results and methods to other populations

Figure 4 Figure 4: The standardized and individualized test protocols to test maximal (an)aerobic capacities on an instrumented ergometer.

Figure 5 Figure 5: Wheelchair mobility performance measurements during the Wheelchair Rugby World Cup in Veijle.

Figure 6 Figure 6: Sensor placement in the wheel for capturing performance data.

will help persons in wheelchairs to become independent, participate in society and stay physically active and healthy. In principle every manual-wheelchair user is an upper-body athlete in need of the best training facilities.

In summary, WheelPower developed tools to measure power output during training and competition and to assess the maximal (an)aerobic capacities in a standardized environment. By adding the power output to the WMPM app, we can quantify the training load of players throughout the season and relate these to the maximal achieved (an)aerobic power output. Performance measures are gathered in a "Performance Portfolio" which is fed back to the trainers, coaches and embedded scientists. The "Performance Portfolio" will be gradually filled with relevant performance facilitators and performance killers so that athletes can optimise their wheelchair settings and/or training guidelines to perform maximally during competition. Furthermore, gathered performance data and research could aim in more evidence informed classification guidelines, to enhance the fairness of wheelchair sports, which is a key topic in Paralympic sports development.

References

1 Cheri Blauwet, Jan Lexell, Wayne Derman, Guzel Idrisova, James Kissick, Jaap Stomphorst, Yetsa Tuakli Wosornu, Peter Van de Vliet, and Nick Webborn. The road to rio: medical and scientific perspectives on the 2016 paralympic games. *PM&R*, 8(8):798–801, 2016.

2 Anthony Bouillod, Julien Pinot, Georges Soto-Romero, William Bertucci, and Frederic Grappe. Validity, sensitivity, reproducibility, and robustness of the powertap, stages, and garmin vector power meters in comparison with the srm device. *International Journal of Sports Physiology and Performance*, 12(8):1023–1030, 2017.

3 Sonja de Groot, Femke Bos, Jorine Koopman, AE Hoekstra, and RJK Vegter. Effect of holding a racket on propulsion technique of wheelchair tennis players. *Scandinavian journal of medicine & science in sports*, 27(9):918–924, 2017.

4 R De Klerk, RJK Vegter, HEJ Veeger, and LHV Van der Woude. a novel servo-driven dual-roller handrim wheelchair ergometer. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(4):953–960, 2020.

5 K Griggs, V Goosey-Tolfrey, and T Paulson. Supporting paralympic wheelchair sport performance through technological, physiological and environmental considerations. *Ann. Hum. Biol.*, 44(4):295–296, June 2017.

6 Liyun Guo, Andrew M Kwarciak, Russell Rodriguez, Nilanjan Sarkar, and W Mark Richter. Validation of a biofeedback system for wheelchair propulsion training. *Rehabilitation Research and Practice*, 2011, 2011.

7 Rowie JF Janssen, Sonja De Groot, Lucas HV Van der Woude, Han Houdijk, and Riemer JK Vegter. Toward a standardized and individualized laboratory-based protocol for wheelchair-specific exercise capacity testing in wheelchair athletes: A scoping review. *American Journal of Physical Medicine & Rehabilitation*, 102(3):261–269, 2023.

8 Rowie JF Janssen, Riemer JK Vegter, Han Houdijk, Lucas HV Van der Woude, and Sonja de Groot. Evaluation of a standardized test protocol to measure wheelchair-specific anaerobic and aerobic exercise capacity in healthy novices on an instrumented roller ergometer. *Plos one*, 17(9):e0274255, 2022.

9 Thomas Paulson and Victoria Goosey-Tolfrey. Current perspectives on profiling and enhancing wheelchair court sport performance. *International Journal of Sports Physiology and Performance*, 12(3):275–286, 2017.

10 Sean D Shimada, Rick N Robertson, Michael L Bonninger, and Rory A Cooper. Kinematic characterization of wheelchair propulsion. *Journal of rehabilitation research and development*, 35(2):210–218, 1998.

11 Erik van der Graaff, Marco MJM Hoozemans, Martijn Nijhoff, Michael Davidson, Merel Hoezen, and Dirkjan HEJ Veeger. Timing of peak pelvis and thorax rotation velocity in baseball pitching. *The Journal of Physical Fitness and Sports Medicine*, 7(5):269–277, 2018.

12 Rienk MA van der Slikke, Monique AM Berger, Daan JJ Bregman, and Dirkjan HEJ Veeger. Wearable wheelchair mobility performance measurement in basketball, rugby, and tennis: lessons for classification and training. *Sensors*, 20(12):3518, 2020.

13 Rienk MA Van der Slikke, Paul Sindall, Victoria L Goosey-Tolfrey, and Barry S Mason. Load and performance monitoring in wheelchair court sports: a narrative review of the use of technology and practical recommendations. *European Journal of Sport Science*, 23(2):189–200, 2023.

14 RMA Van der Slikke, MAM Berger, DJJ Bregman, AH Lagerberg, and HEJ Veeger. Opportunities for measuring wheelchair kinematics in match settings; reliability of a three inertial sensor configuration. *Journal of Biomechanics*, 48(12):3398–3405, 2015.

15 Marit P van Dijk, Manon Kok, Monique AM Berger, Marco JM Hoozemans, and DirkJan HEJ Veeger. Machine learning to improve orientation estimation in sports situations challenging for inertial sensor use. *Frontiers in Sports and Active Living*, 3:670263, 2021.

16 HEJ Veeger and FCT Van Der Helm. Shoulder function: the perfect compromise between mobility and stability. *Journal of biomechanics*, 40(10):2119–2129, 2007.

17 Riemer JK Vegter, Claudine J Lamoth, Sonja De Groot, Dirkjan HEJ Veeger, and Lucas HV Van der Woude. Variability in bimanual wheelchair propulsion: consistency of two instrumented wheels during handrim wheelchair propulsion on a motor driven treadmill. *Journal of NeuroEngineering and Rehabilitation*, 10:1–12, 2013.

18 Chris Whittle, Neal Smith, and Simon A Jobson. Validity of powertap p1 pedals during laboratory-based cycling time trial performance. *Sports*, 6(3):92, 2018.

Improving Power System Resilience with Enhanced Monitoring, Control, and Protection Algorithms

Nidarshan Veerakumar
Faculty of EEMCS, Delft University of Technology, The Netherlands

Aleksandar Boričić
Faculty of EEMCS, Delft University of Technology, The Netherlands

Ilya Tyuryukanov[1]
Faculty of EEMCS, Delft University of Technology, The Netherlands

Marko Tealane
Elering AS, Tallinn, Estonia

Matija Naglič
TenneT TSO B.V., Arnhem, The Netherlands

Maarten Van Riet
Alliander DSO B.V., Arnhem, The Netherlands

Danny Klaar
TenneT TSO B.V., Arnhem, The Netherlands

Arjen Jongepier
Stedin B.V., Rotterdam, The Netherlands

Jorrit Bos
TenneT TSO B.V., Arnhem, The Netherlands

Mohammad Golshani
General Electric Digital, Edinborough, UK

Gert Rietveld
VSL / University of Twente, Enschede, The Netherlands

Mart van der Meijden
TenneT TSO B.V., Arnhem, The Netherlands
Delft University of Technology, The Netherlands

Marjan Popov[2]
Faculty of EEMCS, Delft University of Technology, The Netherlands

Abstract

This paper deals with the essentials of synchrophasor's applications for future power systems to increase system reliability and resilience, which have been investigated within a four-year research project. The project has several applications, covering real-time disturbance detection and blackout prevention distributed across multiple work-packages. Firstly, an advanced big-data management platform built in a real-time digital simulation (RTDS) environment is described to support measurement data collection, processing, and sharing among stakeholders. This platform further presents and demonstrates a network-splitting methodology to avoid cascading failures. Online generator coherency identification is another synchrophasor application implemented on the platform, the use of which is demonstrated in the context of controlled network splitting. Using synchrophasors, data-analytics techniques can also identify and classify disturbances in real time with minor human intervention. Therefore, a novel centralized artificial intelligence (AI) based expert system is outlined to detect and classify critical events. Finally, the paper elaborates on developing advanced system resilience metrics for real-time vulnerability assessment of power systems with a high penetration of renewable energy, focusing on increasingly relevant dynamic interactions and system instability risks.

2012 ACM Subject Classification Hardware → Power and energy; Theory of computation → Design and analysis of algorithms

Keywords and phrases Grid Resilience, Synchrophasors, Real-time Cyber-Physical Experimental Testbed, Real-Time Monitoring, Protection, and Control, Event Detection Classification, Artificial Intelligence, Adaptive Incremental Learning, Controlled Islanding, Vulnerability, State Estimation, Dynamic Line and Cable Rating

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.7

[1] Since October 1, 2021, Dr. Ir. Ilya Tyuryukanov is with Siemens Energy AG, 91058 Erlangen, Germany.
[2] The corresponding author.

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 7; pp. 7:1–7:18
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Funding This work was financially supported by the Dutch Scientific Council NWO in collaboration with Transmission System Operators (TSO) TenneT, Distribution System Operators (DSOs) Alliander, Stedin, VSL and General Electric in the framework of the Energy System Integration and Big Data program under the project "Resilient Synchromeasurement-based Grid Protection Platform, no. 647.003.004".

1 Introduction

Electrical Power Systems (EPSs) undergo significant changes which result from the high penetration of renewable resources. The increased amount of renewable energy shows a negative impact on system inertia and strength. Additionally, using more power electronics causes reduced and bi-directional fault currents, which are difficult to detect. Another significant aspect is the continuous aging of the equipment (especially cables), which may sometimes fail due to old insulation and cause unwanted tripping of the power system protection (e.g. Amsterdam blackout, dated January 17, 2017). All these matters raise questions: Will the existing monitoring, protection, and control schemes be sufficient to cope with new system dynamics imposed by intermittent renewable resources? How can the system be more secure and more resilient to new phenomena (low fault currents, inertia, harmonics, power swings, and protection maloperation)? These questions are crucial for system operators, transmission and distribution utilities, as well as suppliers of power equipment. The ongoing energy transition requires establishing new platforms for monitoring, protection, and control, which will be essential to increase the reliability and security of supply. This is the main goal of this research project, supported by a broad consortium of industrial partners.

Synchronized measurement technology (SMT) utilizes Phasor Measurement Units (PMU) [22] to deliver time-synchronized wide-area measurements near real-time. The SMT is the key building block of the Wide Area Monitoring, Protection, Automation and Control (WAMPAC) system [27, 32], which can improve system security, stability, and reliability. The conventional PMU measurements are obtained at 50/60 frames per second, which provides EPS dynamic behavior observability within the 10 Hz range. Additionally, point-on-wave measurements enabled by powerful SAS sensors technology available in the Netherlands can observe frequencies of up to 4 kHz [4]. In the future, novel monitoring platforms for EPS will be needed to merge PMU and IEC 61850 Sampled Values measurement data. At the same time, providing feedback to the system based on decision-making algorithms depending on the types of disturbance. IEEE Std 1159 [2] defines and summarizes classification methods, and IEC 61000-4-30 [33] defines classes of disturbances. These standards define the classification based on the waveform characteristics (like switching overvoltage, fault current, and voltage dips). The key contributions of this project are: i) a novel synchronized data management platform is presented, which collects measurements in real time with advanced post-processing techniques. The purpose of this platform is to host integrated novel protection and control schemes and other applications derived from real-time updated datasets. ii) A new out-of-step protection algorithm based on real-time synchrophasor measurement data has been developed and compared to existing solutions. The new solution is shown to be more effective and faster than the existing solutions, realized by using the latest General Electric equipment. iii) Developing artificial intelligence models based on incremental learning to detect and classify specific disturbances. These models can be applied to detect anomalies that may arise in the power system due to equipment aging and, in this way, to prevent sudden component failure. iv) Finally, short-term voltage stability and grid vulnerability assessment in the transmission and distribution grids have been investigated to evaluate system strength and the risk of cascading in order to prevent system voltage collapse.

The paper is organized as follows: Section II provides a general overview of the problem and the methodology of its solution. This section also demonstrates the potential application perspective and elaborates on developed concepts. Section III briefly elaborates on the project consortium structure and the individual contributions of its members. Section IV deals with the key results and addresses each project work package separately. In this way, the new platform is shown to enable the detection of both electromechanical and electromagnetic disturbances. Finally, the paper ends by describing the future possible steps and prospects.

2 Project description and motivation

The ReSident[1] project is motivated by the challenges of controlling future electric power grids, such as reduced system resilience and increased operational uncertainty due to the large-scale integration of renewable energy sources. By using advanced tools and algorithms and a comprehensive real-time Energy Management System (EMS), the classification of disturbances and assessment of their impact is aimed to be realized to take preventive operating actions and real-time remedial actions for anticipating and preventing significant system failures. The visualized overview of the ReSident project is given in Figure 1.

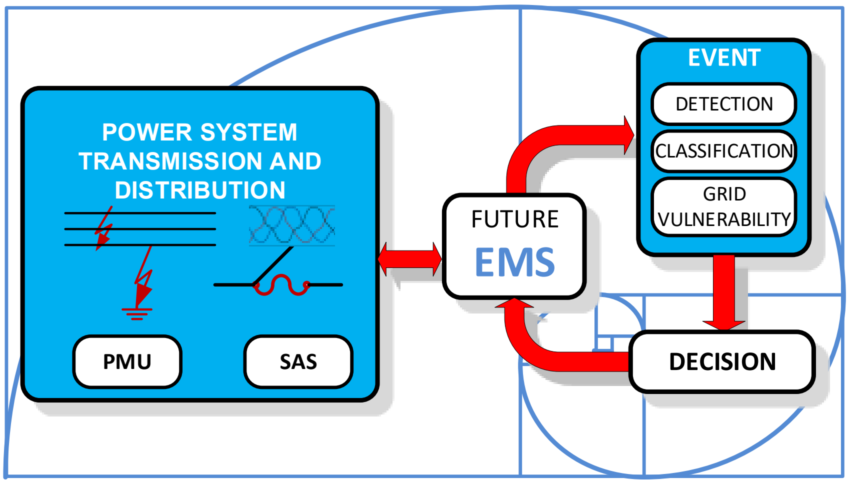


Figure 1 Resident project overview.

The research scope of the Resident project consists of four work packages (WPs). The WPs share a common goal of addressing the resilience of modern power systems from various perspectives. This is highlighted in Figure 1, including the illustrated relation between different tasks. The four WPs and their respective research objectives are listed below.

- ***WP1–Next-generation Energy Management System (EMS) Platform:*** The first work package sets up the cyber-physical experimental test bed. It deals with creating a real-time communication platform for the simulation of power systems based on real-time data, as well as testing and validating WAMPAC applications. Upgrading the SCADA-based telecommunication system with SMT enables greater observability and enhances real-time contingency analysis capabilities. The first step is to estimate the system states (nodal voltage magnitudes and angles) and achieve high real-time grid

[1] Resilient Synchromeasurement-based Grid Protection Platform.

observability. Vast knowledge of network topology, breakers status information, line/cable impedances, and other parameters previously collected for offline load flows are used to create a model-based real-time situational awareness platform. Furthermore, this platform is enhanced by the additional measurements received by the online dynamic line rating (DLR) tool, which manages network congestion that arises due to a highly intermittent RES-dominated power grid.

- ***WP2–Event Detection, Localization and Classification:***
  This work package continues on WP1 and focuses on developing WAMPAC applications that improve real-time situational awareness and perform online pre-contingency analysis. In this context, the first task is to identify disturbances and locate the sub-areas and components responsible for these disturbance events. The second task deals with data-mining to check when the event has been foreseen. Third, when the event has already been foreseen, the event type is classified in real-time. Fourth, when the event is novel and arises from newly installed technologies, dynamic incremental learning is used to update the AI model in near-real time. This ensures an AI-based event classification model adapting through grid transition, supporting network operators with a pre-contingency analysis.
- ***WP3–Stability-aware Controlled Network Separation:***
  This work package focuses on deriving methods for the early detection and prevention of out-of-step (OOS) system conditions. This is facilitated by using wide-area measurement information and developing a controlled system separation and re-stabilization scheme. The first task is to tackle controlled network separation, and instability has to be recognized. After a disturbance occurs, the risk for loss of system stability is monitored concerning several machine-related and system-wide stability indices, and coherent generator groups are identified. The optimal splitting cutsets and control actions are developed and tested on the WP1 experimental test bed, based on which the controlled network separation is executed. Additionally, advanced OOS protection schemes are developed, comprehensively tested, and implemented in the field.
- ***WP4–Grid Vulnerability and Cascading Failures Prevention:***
  The last work package deals with developing offline and online algorithms and tools for the vulnerability assessment of modern power grids. An unprecedented evolution occurs in electric power systems due to the integration of massive amounts of renewable generation, leading to novel and complex stability challenges. Concurrently, robust conventional fossil-based generation is phased out, resulting in resilience challenges. Therefore, it becomes increasingly important to understand and accurately evaluate the vulnerability and resilience of modern power systems so that the risk of instabilities occurring (e.g. uncontrolled separation from WP3) can be minimized. This work package aims to develop algorithms to quantify, anticipate, describe, and prevent cascading failures and system collapse risks by utilizing both offline as well as the real-time data platform from WP1.

The detailed work of all four WPs is described further in Section 4.

3 Project consortium structure and cooperation

The consortium of this large project consists of experts with different backgrounds related to transmission and distribution system operation, measurements, and wide area monitoring and protection. TSO TenneT, and DSOs Stedin and Alliander are utilities that facilitated the project by providing significant data for the transmission and distribution grids that were examined and based on which the algorithms were developed and tested. One of the grids that was taken as an example is a distribution grid operated by Stedin, which has been modeled in a real-time digital simulator. As a manufacturer and one of the main suppliers of equipment for Dutch utilities, General Electric (GE) provided equipment and training for the researchers who verified their algorithms by commercial GE Relays. One of the developed applications has also been installed in the Icelandic electrical power system to detect islanding conditions and provide relevant protection. Another application of GE relay was used for hardware-in-the-loop validation of the dynamic line rating tool for real-time ampacity calculations using solar and wind weather data provided by VSL. Furthermore, the consortium significantly contributed to various engaging technical discussions throughout the project duration that helped shape and improve the developed solutions.

4 Key results

This section discusses the key findings from various WAMPAC applications, which can be congregated into three groups based on the challenges addressed. Applications belonging to the first group aim to achieve real-time situational awareness and improve power system dynamic performance in the long run. Typical examples are real-time state estimation for comprehensive EPS observability and real-time dynamic line rating for online identification of network congestion described in WP1. The second group in WP2 forms a set of pre-contingency-related WAMPAC applications for detecting and mitigating disturbances arising from planned or unplanned power system events, such as real-time disturbance detection, localization, and classification. Further, with the increasing interest in data-driven approaches, we discuss the importance of online training in data-based detection and classification models. Online and near-real-time model training improves parametric sensitivity and avoids misclassification due to concept and/or data drift. Finally, in WP3 and WP4, the third group consists of applications tailored for post-contingency to reduce the adverse impact of the events and improve the system restoration through centralized remedial action schemes. Typical examples are controlled islanding and post-event grid vulnerability assessment.

4.1 WP1–Next-generation Energy Management System (EMS) Platform

Power system control paradigms are shifting from the traditional local control schemes to the modern wide-area synchrophasor data-assisted control schemes. The next-generation EMS platform constitutes a collection of computerized workstations responsible for maintaining the stability of inter-operating AC-HVDC-Smartgrid infrastructure. The platform should also guarantee a secure, reliable, and economically efficient energy management system [20]. In this context, the SMT-supported and WAMPAC-ready cyber-physical testbed is developed as shown in Figure 2. This testbed constitutes 4 layers. From the bottom up, the physical system layer runs simulations to characterize power system dynamics. The networked physical component layer models synchrophasor measurement devices, smart meters, and relays that are equipped with telemetry and, simultaneously, interact with the physical system. The communication network emulation layer creates software architecture for the data transfer. The application layer is where to host various synchrophasor applications. The web-based SMT monitoring platform named Synchro-measurement Application Development Framework (SADF) fills the scientific gap between IEEE Std. C37.118.2-2011 specifications and implementation support for the online synchro-measurement data collection. SADF software library, for the first time, enabled online receiving and parsing of the machine-readable format in the MATLAB programming environment in real-time. Also, we highlight that SADF is the first open-source, available implementation for the IEEE Std. C37.118.2-2011 [22]. SADF is extended further with parallel processing architecture instead of serial processing to utilize high reporting-rate streaming PMU data for processing complex, computationally intensive model-based and data-based (AI) algorithms.

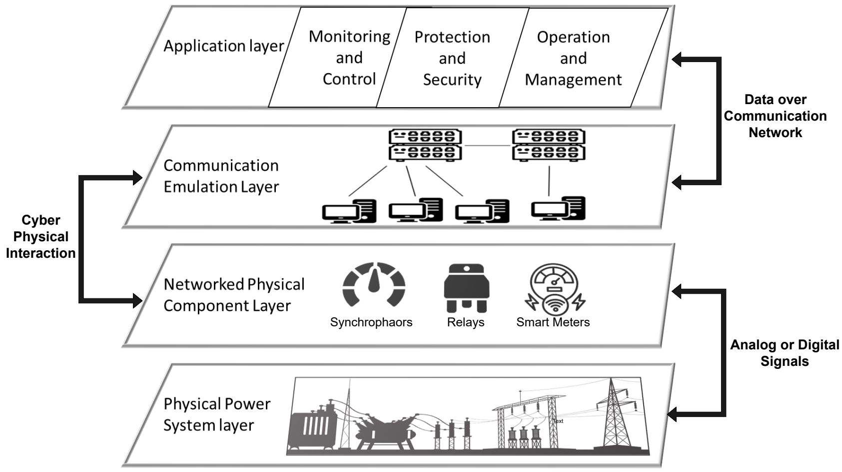


Figure 2 SMT-supported WAMPAC-ready cyber-physical testbed.

As a first application, we utilize our test-bed to realize the true operational state of the grid which is the goal of any TSO or DSO. With streaming time-synchronized PMU data, it is now possible to transit from industrial standard SCADA-powered static state estimation to real-time PMU-based dynamic state estimation [31]. To demonstrate this, a real-time PMU-based state estimation platform for distribution grids is developed, tested, and validated using RTDS. The developed platform serves as proof of concept for potential implementation in an existing 50 kV ring network of the Dutch distribution utility Stedin located in the southwest of the Netherlands (Zeeland province). To catch up with the fast sampling rates of PMUs, the platform incorporates computationally efficient techniques for state estimation. We also incorporate advanced detection, discrimination, and identification of anomalies like bad data (BD) and sudden generation and load changes. Forecasting Aided State Estimation (FASE) has been utilized to enable measurement innovations needed for fast anomaly detection, discrimination, and identification (ADDI), whilst the Extended Kalman Filter (EKF) algorithm is selected to provide fast state forecasting and filtering. The platform has been tested under various normal and abnormal operating conditions considering different statistical properties of sudden load change, measurement noise, and BD scenarios. To demonstrate the advantages and disadvantages of embedding EKF into the platform, EKF is compared with Unscented Kalman Filter (UKF) regarding estimation accuracy, computational efficiency, and compatibility with the module for ADDI. To our knowledge, this was the first validation experiment of the potential implementation of PMU-based real-time state estimation for a real-life distribution grid, achieved through an SMT-supported WAMPAC-ready cyber-physical testbed.

The next WAMPAC application deals with real-time dynamic line rating (RT-DLR). Often, system operators face network congestion issues forcing them to implement preventive or corrective measures, such as generation rescheduling (re-dispatch), with undesirable economic consequences for the system operator [24]. The foreseeable network congestion achieved through state estimation originates from a coincidence of several factors: 1) High reliance on electrical energy has resulted in the increased electrification of residential, transportation, and industrial energy demand, 2) Large costs and difficulties in constructing new overhead transmission lines and underground cabling assets, 3) Significant growth in the share of intermittent Renewable Energy Sources (RES).

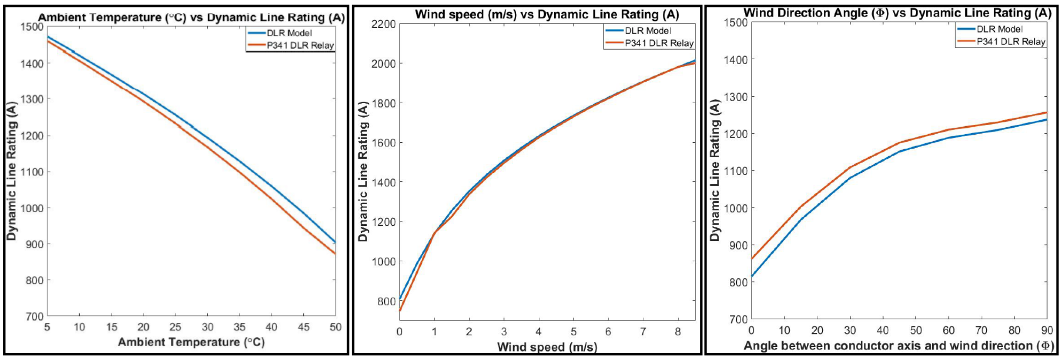


Figure 3 Comparative analysis between RT-DLR and P341 DLR Relay on DLR calculated for the variation in ambient temperature, wind speed, and wind direction.

RT-DLR uses a software-in-the-loop (SIL) test setup consisting of three main parts. The first is the same electrical network utilized by the real-time PMU-based dynamic state estimation, followed by the second part, the weather model. An API is utilized to obtain updated weather parameters, such as ambient temperature, wind speed, and latitude, from the weather stations deployed at the selected location into the MATLAB programming environment. The last part is the resistance and ampacity calculation. The line resistance is calculated using a PMU-based algorithm using bus voltages and branch currents. Considering the weather parameters' influence, a SIL operation by updating the RTDS network is created. The developed DLR model is validated against GE's MicoM Agile P341 DLR relay through a Hardware-in-the-loop setup. Figure 3 describes a comparative analysis between RT-DLR and P341 DLR Relay on ampacity calculated for the variation in ambient temperature, wind speed, and wind direction. It is noted that our RT-DLR-based ampacity calculation method emulates the results of the GE relay under different weather conditions.

4.2 WP2–Disturbance Detection, Localization, Classification, and Learning

With the advanced communication and sensory protocols outlined under IEC61850 SV and IEEE C37.118.2 standards, network operators are subjected to receive streaming high-sampled data in real time. This advancement is raising interest in data-centric models due to: 1) the abundant availability of high-quality data. 2) proven success in parallel disciplines of artificial intelligence. One of the revolutionary features of the WAMPAC platform is its capability to offer a high level of situational awareness and dependable grid operations. To this end, the advancement of real-time detection, localization, classification, and learning (DLCL) of disturbance events draws interest from the scientific community as it improves centralized control and remedial action schemes under decentralized low inertial grid conditions.

The work of disturbance DLCL has four noticeable components. The first component is disturbance detection which picks up any abnormality in the PMU streams deviating from the quasi-steady-state operation. Disturbance detection in steady state conditions is increasingly becoming complex for the following reasons: 1) Reduced inertia due to decommissioning large rotating generators has made power systems vulnerable to small dynamic changes; 2) high-

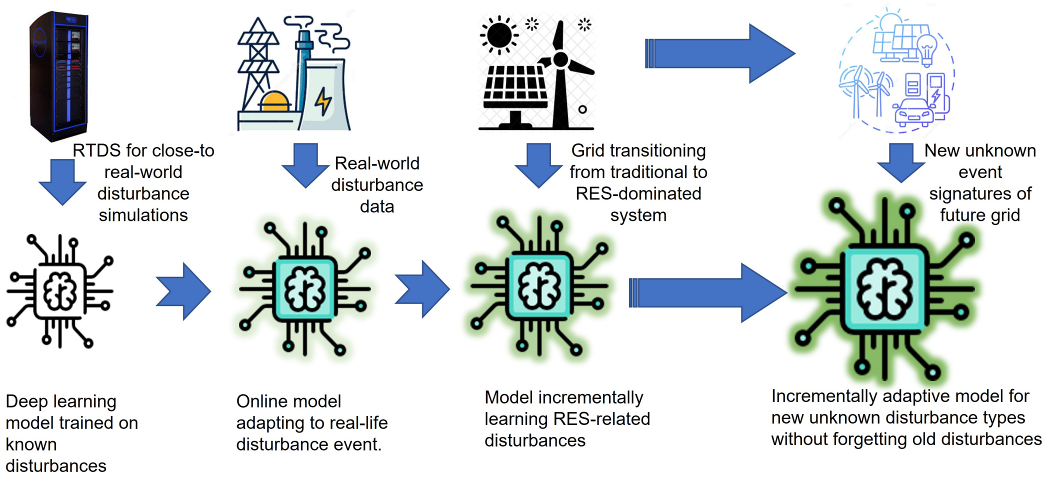


Figure 4 Illustration of an adaptive incremental (AI) model training for disturbance classification.

frequency switching devices in RES and large loads inject harmonics; 3) Multi-dimensional agents controlling the grid operations is incurring sudden generation and load changes. The PMU streams acquired from such sources will be dynamic, noisy, and unpredictable. Further, detecting disturbances upon such PMU streams will be challenging for any non-adaptive statistically-based event detection algorithm [11, 23, 16]. The state-of-the-art has presented many model-based and data-based detection algorithms. Data-based algorithms are gaining more popularity against model-based approaches (digital signal processing and non-adaptive statistical techniques) due to abundant data availability and online training capabilities. These approaches can be analyzed using relative risks and perceived opportunities. In this research, three robust adaptive statistical disturbance detection schemes are developed and analyzed and a data-based real-time adaptive self-learning disturbance detection scheme is proposed [31, 12, 21]. The algorithms are tested under stringent AC and HVDC grid conditions with varying noise, monitoring window length, and several PMU streaming devices.

The second component comprises event localization in a power system. Even though data-based algorithms can identify the PMU devices that are picking up the disturbances, physics-backed model-based approaches based on severity index are more accurate in locating the exact event location. We proposed a method to successfully detect the faulted transmission line, fault type, and the distance to the fault [13]. With our SADF-based RTDS-Matlab co-simulation platform, we developed a PMU-voltage drop-based fault locator, which splits the grid into a user-determined number of subareas based on the PMU locations. Thereafter, an in-depth search is carried out on the faulted area to determine the faulted line. Finally, using a distributed parameter model, fault distance could be determined.

The third and fourth component discusses real-time event classification and learning. The event classification algorithm identifies the event type causing the disturbance signatures. Recent scientific literature has shown great confidence in AI-based classification models, as they can classify events with close to 100% accuracy. Unlike disturbance detection, classification models are not prone to network scalability since only a limited number of IED devices pick up the disturbances at any given time; however, functional scalability of the classification algorithm is a challenging task with a growing set of new event types due to grid transition. Figure 4 envisions training an AI model incrementally for event classification. Just like the human thought process, the algorithm should be able to recognize and discriminate a disturbance event as a new event class or an event from a pre-trained event class. This part considers an assumption in most disturbance event classification literature [30].

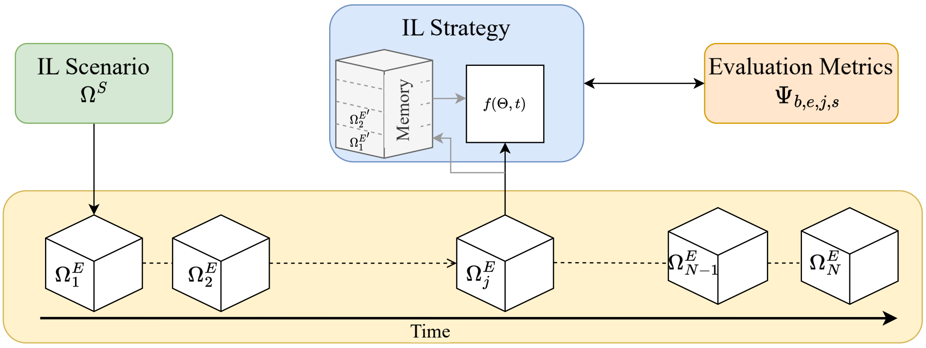


Figure 5 Description of incremental learning strategy - training a deep learning model with new event type scenarios further controlled by evaluation metrics.

Furthermore, if the disturbance encountered is new, the model will be trained incrementally for the new event type on top of the previously trained model. One issue with repeated training of new event types is that the model forgets the intelligence built to remember the previously trained events. This may lead to failure to identify similar disturbances of the same type in the future. This forgetting caused by repeated memory overwriting, is called catastrophic forgetting [18]. The proposed dynamic incremental learning method in Figure 5 addresses this catastrophic forgetting phenomenon to safely update data-based event classification models deployed in future control rooms[30].

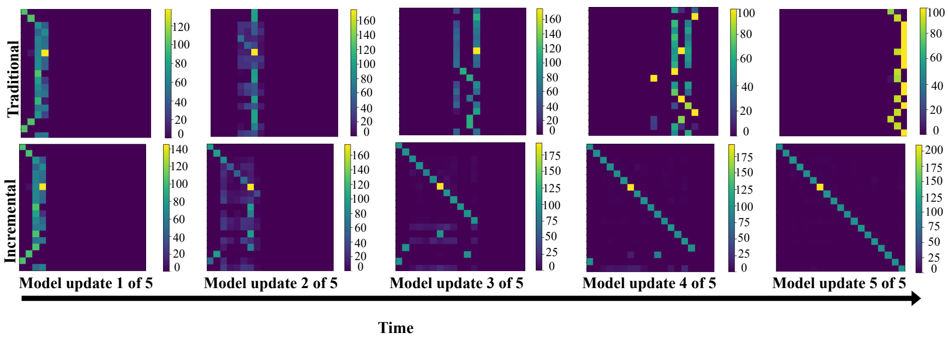


Figure 6 Comparative analysis between online training of deep learning model using a traditional strategy and incremental learning strategy.

By using evolving confusion matrices Figure 6 provides a comparative analysis between online training of deep learning model using traditional and incremental learning strategies. The first row in Figure 6 depicts how a model tends to forget previously trained event types when updated online on the fly. This can be seen from the vertical bands that shift to a new position after each model update. However, in the second row, the strong alignment of the diagonal elements indicates high classification accuracy for each model update. The method is designed to learn efficiently for incoming disturbance data with minimized training time and the highest classification accuracy eliminating catastrophic forgetting. The obtained results show the methodology's performance based on classification accuracy, training time, and storage memory.

4.3 WP3–Stability-aware Controlled Network Separation

Modern interconnected electric power systems (EPS) face significant challenges due to several factors, such as electricity market deregulation, the increased utilization of intermittent renewable energy sources, and aging equipment. As a result of the confluence of these and other factors, there has been a rise in the incidence of severe EPS events and blackouts in recent years [14]. Noteworthy examples of such events include splitting the European power grid in January and July 2021, the 2021 Texas blackouts, and the 2016 South Australian blackout. Generally, large-scale power failures are triggered by an uncontrolled sequence of events that culminate in system instability. Figure 7 provides a schematic representation of a typical blackout timeline [17].

Figure 7 Typical sequence of events in a blackout scenario.

The operating condition of the EPS typically undergoes a gradual decline before experiencing a failure, as depicted by the steady-state progression in Figure 7. This decline is often attributed to an unforeseen triggering event in an otherwise healthy power system. Once the EPS enters a vulnerable state, a single triggering event can lead to instability or prompt rapid cascading outages that culminate in instability and blackout. The controlled system separation approach is intended to counteract this chain of events by detecting instability early and immediately computing the optimal line cutset. By doing so, the unstable region of the EPS is separated while minimizing the loss of load and equipment overloads.

The process of controlled separation involves three main decisions: *when to split*, *where to split*, and *what to do after splitting*. These decisions also represent the different stages of the network splitting algorithm. The *when to split* decision is responsible for identifying stable and unstable transients in the EPS and detecting the key EPS elements that cause instability, as this set of elements may change with time. The *where to split* decision involves choosing the optimal EPS branches to disconnect to isolate the unstable part of the EPS. Finally, the *what to do after splitting* decision involves calculating corrective control actions such as load shedding to stabilize each island formed after splitting in a cost-efficient manner.

The first stage of the controlled separation problem is focused on instability recognition. The scope is limited to detecting loss of synchronism (LoS), the most important instability type requiring system separation. LoS in an EPS typically occurs along the boundaries of coherent generator groups (CGG), which are groups of generators that tend to swing together following a disturbance. CGGs are identified using the novel method defined in [21] or operational experience/measurement-based coherency identification [29]. After disturbance detection, the EPS is monitored concerning several machine-related and system-wide stability indices to track the risk of LoS. When the risk of LoS becomes very high, the system split command is activated, and the most disturbed CGG is separated from the grid. This decision logic of system splitting is summarized in Figure 8.

For each CGG, the control actions to stabilize the islands are continuously recomputed based on recent snapshots of the system state. Finding the optimal control actions is formulated as a linear or non-linear integer program to satisfy more physical power system constraints. However, measurements during intensive power swings cannot easily characterize the network's power balance and loading conditions. Thus, the computational time delay between measurement and control is partly compensated by the duration of power swings

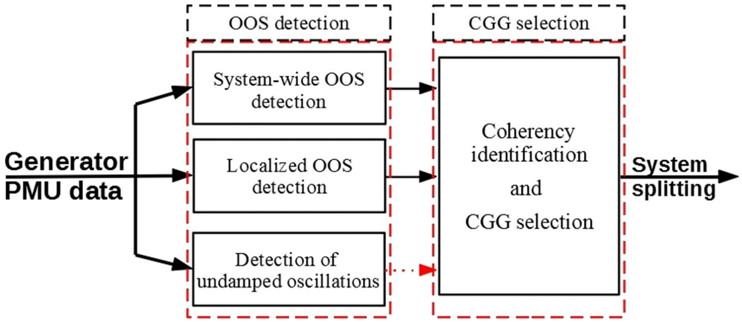


Figure 8 Decision logic of controlled splitting.

before instability detection, during which no viable state estimates can be obtained. Once the first stage detects instability and the most disturbed CGG, the most recent splitting solution computed for that CGG is activated to separate it from the grid in a stable and cost-efficient manner. An example of the controlled splitting process is hereby illustrated, based on an IEEE 39 bus network described in [21]. The tripping of an important line causes the generator connected to bus 38 to experience growing oscillations, eventually becoming out-of-step with the rest of the system. This is depicted in Figure 9.In the left part of Figure 9, the top plot illustrates the angles of generator terminal voltages in the system center of inertia (COI) frame of reference denoted as Φ_G^{coi}, and the lower plot shows the corresponding frequencies denoted as f_G^{coi}. Both quantities can be measured in real time by PMUs. The coherency estimation algorithm [28] identifies three CGGs, which include generators located at buses 30–37, 38, and 39, displayed in blue, red, and green, respectively.

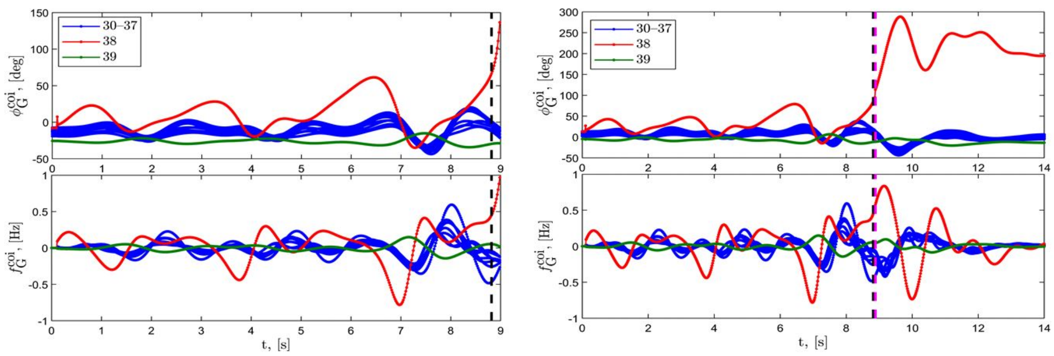


Figure 9 Unstable transient (left) and splitting transient (right).

As depicted in the left part of Figure 9, the generator at bus 38 demonstrates increasing oscillations. Its instability becomes evident 8.7 seconds after the fault inception time, denoted by the black dashed line. The following 0.2 seconds of the sampled transient data verify the instability. Given that the CGG consisting of the generator at bus 38 is the most impacted (displaying the highest average RMS frequency value), it needs to be separated from the rest of the system. The right example of Figure 9 indicates the resulting transient, separating the generator at bus 38 results in two stable islands, thus preventing emerging instability.

Another approach for detecting network instability is to reduce the transmission system into a two-machine system around the observable tie-lines. The impedance of the power network at the terminals of the observed line can be computed by utilizing real-time measurements. After that, using the system impedances, current, and voltage measurements, a power-angle characteristic can be built, and the system's stability can be assessed. This way, an out-of-step (OOS) protection algorithm is developed in a real-time environment [26]. The performance has been compared to the commercially available OOS protection solutions by using RTDS and hardware-in-the-loop methodology. The results confirm that the proposed algorithm detects OOS conditions faster and more reliably than the traditional solutions. However, the reduction of the network has the drawback of requiring a step change to occur in the network to compute the equivalent system impedances. An enhanced algorithm has been developed to overcome this limitation, which is decoupled from power system parameters. The decoupled OOS algorithm utilizes angle difference derivative values to detect instability [26, 25]. The decoupled algorithm has been developed for a hardware platform that receives wide-area measurement system signals.

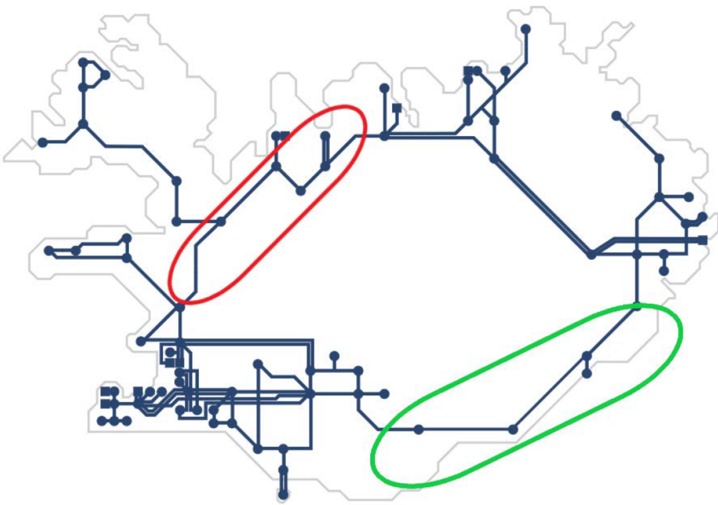


Figure 10 The power system of Iceland, where the newly developed out-of-step algorithm is tested and implemented on the highlighted grid sections.

The performance of the decoupled algorithm has been compared to two existing OOS protection devices available on the market using RTDS hardware-in-the-loop testing, demonstrating faster and more reliable operation. As a result, the decoupled protection algorithm is commissioned and installed in the Icelandic transmission system, illustrated in Figure 10.

4.4 WP4–Grid Vulnerability and Cascading Failures Prevention

In modern power systems that rely on Renewable Energy Sources (RES), technical fundamentals of system operation, security, and resilience are very different from those of conventional power systems. The differences arise primarily due to how RES are integrated into the power system compared to conventional synchronous generators. Synchronous generators (SGs) are massive turbine-powered machines spinning at hundreds to thousands of revolutions per minute. By a carefully designed combination of the rotation speed and spatially distributed energized copper windings, a massive amount of mechanical energy applied by water or steam on the turbine is transformed into electromagnetic force and, ultimately, electrical energy.

At a high level, two main parameters describe the operation of an SG and, consequently, the power system. Those two parameters are *frequency* and *voltage*. The system frequency is proportional to the rotational speed of SGs, in a profound balancing act of supply and demand of active power. When a disturbance occurs, a large amount of rotational kinetic energy stored in the rotating masses is quickly dissipated in a natural attempt to counter the external force attempting to change the state of motion. This is known as *inertia* and is a fundamental concept of conventional power systems. The other aspect of robustness provided by SGs is related to voltage and is a bit more intricate: *system strength*. System strength is a generator's natural ability to counter voltage changes. This effect, however, has nothing to do with rotational mass, but is electromagnetic by nature and is related to the opposition to the change of the magnetic flux between the stator and the rotor of a generator.

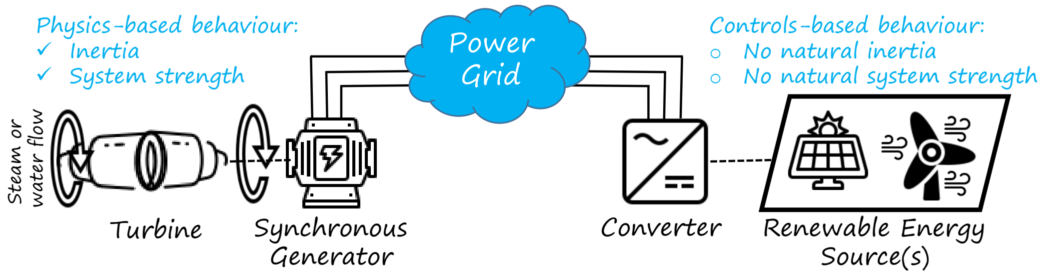


Figure 11 Illustration of some of the fundamental differences between synchronous and inverter-based generation.

Unlike SGs, RES does not have massive rotating machinery with stored kinetic energy or strong electromagnetic coupling to the grid. Instead, they are mechanically and electromagnetically decoupled from the system and are integrated through converters. This means RES, also often referred to as inverter-based resources (IBRs), interact with the system through fast-operating power electronics components, which makes their response more control-based rather than physics-based. Therefore, the behavior of RES is fundamentally different in both steady and (particularly) dynamic state operations, as illustrated in Figure 11.

As the number of RES (SGs) in the grid increases (decreases), the natural resilience of power systems drops. Concurrently, the fast nature of RES controls results in complex dynamics in the grid, which may threaten system stability and increase grid vulnerability. To preserve system resilience, it is necessary to accurately evaluate system strength across both steady and dynamic state operation of power systems [7].

To connect any generation to the grid, one must ensure that the point of connection (PoC) is strong enough for such a connection. This strength is typically evaluated from the perspective of *Short-Circuit Power*, S_{sc} [1]. However, as systems evolve and renewables continuously replace synchronous generation, the concept of S_{sc} loses its significance. To overcome this issue, more accurate methods for system strength evaluation are developed in this project [8]. These more accurately evaluate system strength in the presence of RES and consider other relevant parameters of the grid connection point, such as the operating voltage, shunt elements, and grid impedance. Some of these effects are highlighted in Figure 12.

Besides the steady state operation, the dynamics of power grids are becoming more complex on all voltage levels [15]. A particularly interesting effect is the impact of Distributed Energy Resources (DER), such as solar panels and smaller wind turbines connected to the distribution grid. As the number of DERs rises, the impact on a bulk power system's stability also becomes larger. Furthermore, system demand also becomes more complex, with an increasing number of motors, converter-interfaced loads, and electronic loads.

Extensive studies conducted in [5] reveal how these affect short-term voltage stability and which operating scenarios may lead to vulnerability as the energy transition continues. Some results are illustrated in Figure 13, where the impact of varying penetration of DER and dynamic load on system stability is explored across various parameters.

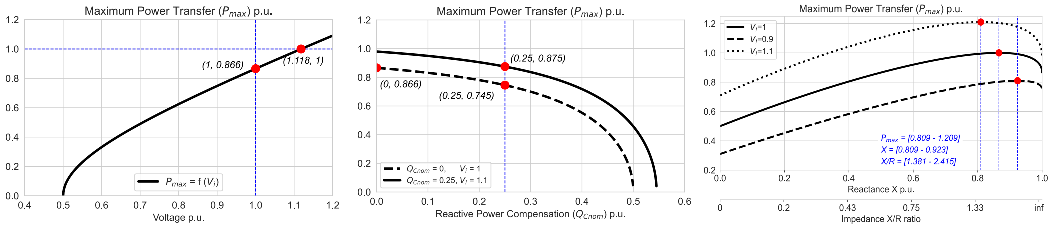


Figure 12 Steady state system strength evaluation in renewable-driven grids with varying parameters: (a) Operating voltage (b) Shunt compensation (c) Varying grid X/R ratio.

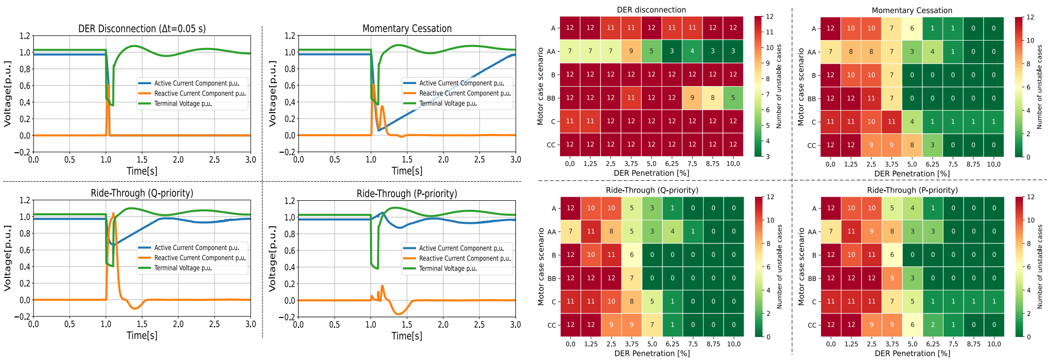


Figure 13 Impact of distributed generation and its controls on grid short-term voltage stability.

As these dynamics and instabilities become more complex, often accompanied by high data uncertainty [19, 3], it becomes increasingly important to utilize and apply advanced data-driven algorithms for stability evaluation [10]. One such algorithm developed within this project is Cumulative Voltage Deviation (CVD) [6, 9], illustrated in Figure 14. The method is designed to automatically utilize either offline simulation data or online measurement data to quantify voltage deviations of various origins. Such an analysis provides dynamic a severity evaluation of potential system disturbances, allowing system operators to proactively react to elevated resilience risks and prevent the occurrence of dangerous cascading faults.

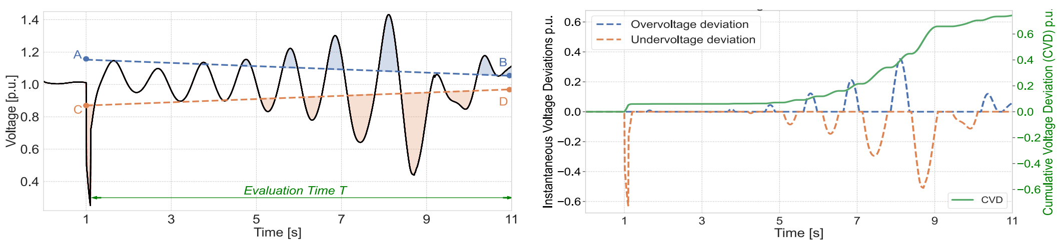


Figure 14 Visualization of the data-based CVD method for quantifying short-term instabilities.

As the resilience of electric power systems decreases, vulnerability analysis becomes crucial. Energy transition reduces inertia and system strength. As a result, faster, intertwined, and more dangerous dynamic phenomena appear. To maintain resilience, evaluation methods also need to evolve. System limitations must be better understood in steady and dynamic states to ensure strength and resilience. As the renewable trends continue, resilience is no longer abundant; instead, it must be (pro)actively provided. Methods developed in this project present a step forward to a more accurate data-based evaluation of system resilience, providing system operators with more information about potential vulnerabilities that pose instability risks and may require operational actions, grid reinforcements, or advanced ancillary services.

5 Further steps and prospects

The Resident project addressed some of the most significant and critical WAMPAC applications for operating modern and future power systems in a more secure and resilient manner. The following prospects are based on our experience with reliable implementation of these applications in future control rooms. A large amount of data availability will require comprehensive hardware to store, analyze and develop models. This means a dedicated workforce of data experts needs to be assigned to TSO's and DSO's who can bridge the gap between data science and vital supervisory planning, operation, and control.

Robust architectural frameworks must be designed to handle applications operating at different time frames. The frameworks should be able to accommodate models that prioritize different performance metrics, such as computational efficiency, accuracy, and memory usage. With the world adapting to data-centric systems for less critical tasks, power system experts and stakeholders should analyze and strategically adapt to changing times. One safe approach is adapting data-centric systems to less critical tasks in network operations, where the stakes of error are lower. Flexible "human-in-the-loop" approaches can be investigated for additional validation of AI-based models. With enough confidence and expertise, hybrid or semi-empirical models can be interwoven with critical workflows such as state estimation, load-flow dispatch, and automatic switching operations corresponding to centralized remedial action schemes. As systems become weaker, they also become more controllable. Using advanced control algorithms and ancillary services to steer systems away from instabilities in real-time across various time scales is a promising field for further research and development.

Additional developments and realization of state-of-the-art tools and algorithms in control rooms and advanced energy management systems are needed to ensure resilient electric power grids. As future power systems must rely on power electronics and increasingly control-driven rather than physics-driven behavior, the additional disturbances and interactions resulting from the interconnection of AC-HVDC electrical grids will inevitably appear more frequently. To cope with these, applying novel real-time synchrophasor-based control and protection schemes will become a necessity for the stable and resilient operation of future power systems.

References

1 Connection of wind farms to weak AC networks. Technical Report WG B4.62, CIGRE, 2016. URL: `https://e-cigre.org/publication/671-connection-of-wind-farms-to-weak-ac-networks`.

2 IEEE Recommended Practice for Monitoring Electric Power Quality. pages 1–98, August 2019. Conference Name: IEEE Std 1159-2019 (Revision of IEEE Std 1159-2009). `doi:10.1109/IEEESTD.2019.8796486`.

3 Guest Editorial: Challenges and New Solutions for Enhancing Ancillary Services and Grid Resiliency in Low Inertia Power Systems. *IET Generation, Transmission & Distribution*, 14(22):4975–4977, 2020. `doi:10.1049/iet-gtd.2020.1491`.

4 Frank Baldinger, Ton Jansen, Maarten van Riet, and Frans Volberda. Nobody knows the future of Smart Grids, therefore separate the essentials in the secondary system. In *10th IET International Conference on Developments in Power System Protection (DPSP 2010). Managing the Change*, pages 1–5, March 2010. `doi:10.1049/cp.2010.0210`.

5 Aleksandar Boričić, Jose Luis Rueda Torres, and Marjan Popov. Fundamental study on the influence of dynamic load and distributed energy resources on power system short-term voltage stability. *International Journal of Electrical Power & Energy Systems*, 131, October 2021. doi:10.1016/j.ijepes.2021.107141.

6 Aleksandar Boričić, Jose Luis Rueda Torres, and Marjan Popov. Quantifying the Severity of Short-term Instability Voltage Deviations. In *2022 International Conference on Smart Energy Systems and Technologies (SEST)*, pages 1–6, September 2022. doi:10.1109/SEST53650.2022.9898503.

7 Aleksandar Boričić, Jose Luis Rueda Torres, and Marjan Popov. System Strength: Classification, Evaluation Methods, and Emerging Challenges in IBR-dominated Grids. In *2022 IEEE PES Innovative Smart Grid Technologies - Asia (ISGT Asia)*, pages 185–189, November 2022. ISSN: 2378-8542. doi:10.1109/ISGTAsia54193.2022.10003499.

8 Aleksandar Boričić, Jose Luis Rueda Torres, and Marjan Popov. Beyond SCR in Weak Grids: Analytical Evaluation of Voltage Stability and Excess System Strength. In *International Conference on Future Energy Solutions (FES2023)*, pages 1–6, June 2023. doi:10.1109/FES57669.2023.10183286.

9 Aleksandar Boričić, Jose Luis Rueda Torres, and Marjan Popov. Quantification and classification of short-term instability voltage deviations. *IEEE Transactions on Industry Applications*, June 2023. doi:10.1109/TIA.2023.3289931.

10 Aleksandar Boričić, José L. Rueda Torres, and Marjan Popov. Comprehensive Review of Short-Term Voltage Stability Evaluation Methods in Modern Power Systems. *Energies*, 14(14):4076, January 2021. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/en14144076.

11 Song Changpo, Bai Jie, Shen Yan, Wang Tianzheng, and Lu Shan. A new detection method of power quality disturbance signal. In J Liu, Y Wang, and H Xu, editors, *Proceedings of the 2015 international conference on intelligent systems research and mechatronics engineering*, volume 121 of *Advances in Intelligent Systems Research*, pages 267–271. Atlantis press, 2015. ISSN: 1951-6851 Type: Proceedings Paper.

12 Jose J. Chavez, Nidarshan Veera Kumar, Sadegh Azizi, Jose L. Guardado, Jose Rueda, Peter Palensky, Vladimir Terzija, and Marjan Popov. PMU-voltage drop based fault locator for transmission backup protection. *Electric Power Systems Research*, 196:107188, July 2021. doi:10.1016/j.epsr.2021.107188.

13 Jose J. Chavez, Nidarshan. Veera Kumar, Marjan. Popov, Peter. Palensky, Sadegh Azizi, Enrique Melgoza, and Vladimir Terzija. Non-Homogeneous Sampling Rate Wide Area Backup Protection using Synchrophasors and IED Data. In *2022 International Conference on Smart Grid Synchronized Measurements and Analytics (SGSMA)*, pages 1–6, May 2022. doi:10.1109/SGSMA51733.2022.9806006.

14 Hassan Haes Alhelou, Mohamad Esmail Hamedani-Golshan, Takawira Cuthbert Njenda, and Pierluigi Siano. A Survey on Power System Blackout and Cascading Events: Research Motivations and Challenges. *Energies*, 12(4):682, January 2019. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. doi:10.3390/en12040682.

15 Nikos Hatziargyriou, Jovica Milanovic, Claudia Rahmann, and et al. Definition and Classification of Power System Stability – Revisited & Extended. *IEEE Transactions on Power Systems*, 36, July 2021. doi:10.1109/TPWRS.2020.3041774.

16 Masoud Karimi, Hossein Mokhtari, and M Reza Iravani. Wavelet based on-line disturbance detection for power quality applications. *IEEE Transactions on Power Delivery*, 15(4), October 2000. doi:10.1109/61.891505.

17 Yuri V. Makarov, Victor I. Reshetov, Andrei. Stroev, and I. Voropai. Blackout Prevention in the United States, Europe, and Russia. *Proceedings of the IEEE*, 93(11):1942–1955, November 2005. doi:10.1109/JPROC.2005.857486.

18 Davide Maltoni and Vincenzo Lomonaco. Continuous learning in single-incremental-task scenarios. *Neural Networks*, 116:56–73, 2019. Publisher: Elsevier. doi:10.1016/j.neunet.2019.03.010.

19 Federico Milano, Florian Dörfler, Gabriela Hug, David J. Hill, and Gregor Verbič. Foundations and Challenges of Low-Inertia Systems (Invited Paper). In *2018 Power Systems Computation Conference (PSCC)*, pages 1–25, June 2018. doi:10.23919/PSCC.2018.8450880.

20 Matija Naglic. *On power system automation: Synchronised measurement technology supported power system situational awareness*. Dissertation (tu delft), Delft University of Technology, 2020. doi:10.4233/uuid:33a9138f-7224-4734-a326-d90a9d5980c1.

21 Matija Naglic, Lian Liu, Ilya Tyuryukanov, Marjan Popov, Mart AMM van der Meijden, and Vladimir Terzija. Synchronized measurement technology supported AC and HVDC online disturbance detection. *Electric Power Systems Research*, 160:308–317, 2018. doi:10.1016/j.epsr.2018.03.007.

22 Matija Naglic, Marjan Popov, Mart AMM van der Meijden, and Vladimir Terzija. Synchro-measurement application development framework: An IEEE standard C37. 118.2-2011 supported MATLAB library. *IEEE Transactions on Instrumentation and Measurement*, 2018.

23 Reza Pourramezan, Houshang Karimi, and Jean Mahseredjian. Real-Time Disturbance Detection and Classification using Principal Component Analysis of PMU Data. In *IEEE Power and Energy Society General Meeting PESGM*. IEEE, 2020.

24 José A. Rosendo-Macías, Antonio Gómez-Expósito, Alfonso Bachiller-Soler, Miguel Á. González-Cagigal, Gabriel Álvarez Cordero, Lucía Mateo-Sánchez, and Antonio Useros-García. The Spanish Experience: Squeezing Line Ampacities Through Dynamic Line Rating. *IEEE Power and Energy Magazine*, 21(1):73–82, January 2023. doi:10.1109/MPE.2022.3219167.

25 Marko Tealane, Jako Kilter, Oleg Bagleybter, Birkir Heimisson, and Marjan Popov. Out-of-Step Protection Based on Discrete Angle Derivatives. *IEEE Access*, 10:78290–78305, 2022. doi:10.1109/ACCESS.2022.3193390.

26 Marko Tealane, Jako Kilter, Marjan Popov, Oleg Bagleybter, and Danny Klaar. Online Detection of Out-of-Step Condition Using PMU-Determined System Impedances. *IEEE Access*, 10:14807–14818, 2022. doi:10.1109/ACCESS.2022.3149103.

27 Vladimir Terzija, Gustavo Valverde, Deyu Cai, Pawel Regulski, Vahid Madani, John Fitch, Srdjan Skok, Miroslav M. Begovic, and Arun Phadke. Wide-Area Monitoring, Protection, and Control of Future Electric Power Networks. *Proceedings of the IEEE*, 99(1), January 2011. doi:10.1109/JPROC.2010.2060450.

28 Ilya Tyuryukanov, Jorrit A. Bos, Mart AMM van der Meijden, Vladimir Terzija, and Marjan Popov. Controlled Power System Separation Using Generator PMU Data and System Kinetic Energy. *IEEE Transactions on Power Delivery*, 2023. doi:10.1109/TPWRD.2023.3249124.

29 Ilya Tyuryukanov, Marjan Popov, Mart A. M. M. van der Meijden, and Vladimir Terzija. Slow Coherency Identification and Power System Dynamic Model Reduction by Using Orthogonal Structure of Electromechanical Eigenvectors. *IEEE Transactions on Power Systems*, 36(2):1482–1492, March 2021. doi:10.1109/TPWRS.2020.3009628.

30 Nidarshan Veerakumar, Jochen L. Cremer, and Marjan Popov. Dynamic Incremental Learning for real-time disturbance event classification. *International Journal of Electrical Power & Energy Systems*, 148:108988, June 2023. doi:10.1016/j.ijepes.2023.108988.

31 Nidarshan Veerakumar, Dragan Ćetenović, Krit Kongurai, Marjan Popov, Arjen Jongepier, and Vladimir Terzija. PMU-based Real-time Distribution System State Estimation Considering Anomaly Detection, Discrimination and Identification. *International Journal of Electrical Power & Energy Systems*, 148:108916, June 2023. doi:10.1016/j.ijepes.2022.108916.

32 Peter Wall, Papiya Dattaray, Zhaoyang Jin, Priyanka Mohapatra, James YU, Douglas Wilson, Karine hay, Stuart Clark, Mark Osborne, Phillip M. Ashton, and Vladimir Terzija. Deployment and demonstration of wide area monitoring system in power system of Great Britain. *Journal of Modern Power Systems and Clean Energy*, 4(3):506–518, July 2016. doi:10.1007/s40565-016-0218-3.

33 Pei Zhang, Fangxing Li, and Navin Bhatt. Next-Generation Monitoring, Analysis, and Control for the Future Smart Control Center. *IEEE Transactions on Smart Grid*, 1(2):186–192, September 2010. doi:10.1109/TSG.2010.2053855.

RATE-Analytics: Next Generation Predictive Analytics for Data-Driven Banking and Insurance

Dennis Collaris
Eindhoven University of Technology, The Netherlands

Mykola Pechenizkiy
Eindhoven University of Technology, The Netherlands

Jarke J. van Wijk
Eindhoven University of Technology, The Netherlands

Abstract

We conducted the RATE-Analytics project: a unique collaboration between Rabobank, Achmea, Tilburg and Eindhoven University. We aimed to develop foundations and techniques for next generation big data analytics. The main challenge of existing approaches is the lack of reliability and trustworthiness: if experts do not trust a model or its predictions they are much less likely to use and rely on that model. Hence, we focused on solutions to bring the human-in-the-loop, enabling the diagnostics and refinement of models, and support in decision making and justification. This chapter zooms in on the part of the project focused on developing explainable and trustworthy machine learning techniques.

2012 ACM Subject Classification Computing methodologies → Philosophical/theoretical foundations of artificial intelligence; Computing methodologies → Machine learning; Human-centered computing

Keywords and phrases Visualization, Visual Analytics, Machine Learning, Interpretability, Explainability, XAI

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.8

Acknowledgements We would like to thank numerous colleagues of the RATE project team at Tilburg University, Rabobank, Achmea, and TU/e who over the years provided endless support and facilitated collaboration. We would like to thank NWO, Rabobank and Achmea for the provided funding. This work is part of the research programme Commit2Data, specifically the RATE Analytics project with project number 628.003.001. Last, but not least we would like to thank the reviewers for providing constructive feedback.

1 Introduction

Banks and insurance companies try to make use of increasingly large volumes of customer-related data as new in-house and external data sources become available for analysis. To seize this opportunity, we conducted the RATE-Analytics project: a unique collaboration between Rabobank, Achmea, Tilburg and Eindhoven University. The overall goal of this project was to develop foundations and techniques for next generation big data analytics, facilitating the development of a wide range of applications in banking and insurance. This includes, but is not limited to: intrusion detection, transactions and insurance fraud detection, loan default prediction, prediction of extreme (insurance) claims, and data-driven product development.

In the algorithms and use cases we studied, the overarching theme is to efficiently combine (large amounts of) data and expert knowledge, where the (economic) cost of the former is rapidly decreasing and the cost of the latter increasing. We, therefore, focus on the expert in the loop, or the expert-algorithm interaction to define what a "good" algorithm is. We bring three major data-science expertise areas into this multidisciplinary project: predictive analytics, modern statistics, and visual analytics.

With this unique combination, we aim at breakthroughs in data-driven banking and insurance, facilitating the development of more reliable, transparent, and responsible analytics solutions and products.

In this chapter, we zoom in on the part of the project focused on developing explainable and trustworthy machine learning techniques. We focus on solutions to bring the human-in-the-loop, enabling the diagnostics and refinement of machine learning models, and support in decision making and justification.

The remainder of this chapter is structured as follows: we describe the problem addressed in Section 2. Next, in Section 3 we describe the consortium structure and way-of-working. The results of the project are highlighted in Section 4 both from a technical, and collaborative perspective. Finally, in Section 5 we conclude by discussing future prospects for the project.

2 Problem addressed

The RATE project addresses the problem of analyzing large volumes of customer-related data in the banking and insurance industries. The various sources of data available at Achmea and Rabobank contain a wealth of information about customer behavior and events. It would be advantageous (academically, for society, and for Achmea/Rabobank) to develop automated methods that take these data and analyze the possible benefits for customers of new finance and insurance products. Concretely, we focus on the problem that existing techniques to achieve this often lack transparency, explainability and trustworthiness. By involving the human-in-the-loop, we can address this and enable the diagnostics and refinement of machine learning models, and support in decision making and justification.

2.1 Challenges

The main challenge of existing data-driven approaches from the application perspective is the lack of reliability and trustworthiness: if experts do not trust a model or its particular output they are much less likely to use and rely on that model. This challenge is not unique for finance and insurance: in many other areas issues such as trustworthiness, explainability, and integration of expert knowledge play an important role. The RATE project provided, from an academic perspective, a unique opportunity to work on these challenges for real-world cases in a huge and complex domain, in close collaboration with the stakeholders.

Without a proper understanding of a machine learning model, there are several problems that can occur in the data science pipeline that may go unnoticed (summarized in Figure 1).

- **ML may be trained on unrepresentative data.** For instance, it could be based on biases, spurious correlations, and false generalizations [16]. As an example, recent work has shown that the accuracy of commercial gender classification on dark-skinned females is significantly worse than on any other group. This discrepancy was found to be largely due to unrepresentative training datasets and imbalanced test benchmarks [4].
- **ML may be using inadequate features.** There may be sufficient and representative data available, but the model could use this data in unexpected ways. Machine learning is only able to ascertain correlation amongst features, and is not able to find causal relationships. This is demonstrated in the wolf-husky problem introduced by Ribeiro et al. [23]. The authors show an example of a husky and wolf classifier, that turns out to detect the type of animal based on snow in the background of an image, instead of looking at the animal itself.

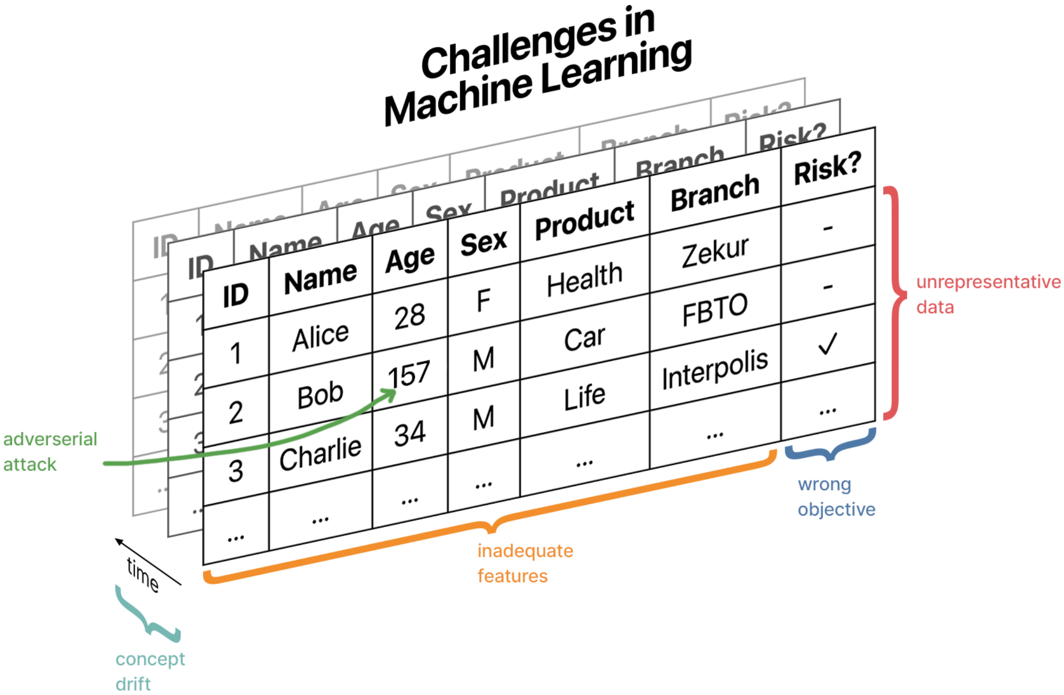


Figure 1 There are several problems that can occur in the data science pipeline that may go unnoticed without a proper understanding of the behavior of a machine learning model.

- **ML may have the wrong objective.** Even if the model scores well on a test set with any of the many performance metrics available (e.g., accuracy, precision, recall, F_1), it may still be trained to optimize the wrong objective, like trying to match doubtful labels. Consider a fraud detection model in insurance. Since no ground truth information is available about which insurance policies are fraudulent, models are typically trained with labels from human fraud experts. In such a scenario, the machine may adopt any biases the fraud experts may have, explicit or implicit. In addition, this model can only be expected to match the human performance, not exceed it.
- **ML may not be robust against concept drift.** The test set generalization during development may not match with future unseen data. This problem was the reason for the failure of the Google Flu Trends model [5], which showed promising results, but failed to predict flu trends in practice. This problem is wide-spread across real-world applications of predictive analytics [26], and is very predominant in adversarial domains (e.g., spam detection, fraud detection).
- **ML may be subject to adversarial attacks.** There is a surge of recent works showing that models may be vulnerable to adversarial attacks. For instance, authors have shown that a small perturbation in the input (e.g., a single pixel in an image or imperceptible noise) can lead to unexpected, extreme changes in the output, often leading to absurd or incorrect predictions [2, 20, 17].

2.2 Application areas

RATE-Analytics aims to address previously mentioned problems by supporting the understanding of complex machine learning models. The stakeholders can use machine learning explanations for a wide variety of applications. We identified four main categories based on discussions with data scientists at Achmea:

1. **Diagnostics** The model may not perform adequately, even though the model scores well on a test set. It could be based on biases and spurious correlations [16]. Explanations can highlight these issues such that experts can address these during model development.
2. **Refinement** Apart from identifying issues with the model, explanations can also help to improve the model. Analyzing explanations for incorrect predictions can yield insights into how to increase predictive accuracy [1, 24] or remove irrelevant features.
3. **Decision support** In high-stakes decision-making, where models make decisions that have a critical impact on real people, it is not sufficient to base decisions on the prediction score of the model alone. To qualitatively ascertain whether desiderata such as fairness, privacy, and trust are met, explanations can help internal decision-makers to verify the behavior of the model [13].
4. **Justification** Various external stakeholders may have questions about predictions by the model. For instance, customers subject to predictions may request justification, or authorities may request information to check compliance. The latter got more relevant since the recent introduction of the GDPR's "right to explanation" [14] and AI Act [15].

The results of the RATE-Anlytics project support data scientists in all these applications through providing a better understanding of machine learning models.

3 Consortium structure

The academic contribution to the project consortium consisted of internationally recognized experts in predictive analytics, visual analytics, and modern statistics who actively participate in national (NWO, STW, COMMIT) and European projects and collaborate with industry. This is a unique combination of three core data science areas in one project. Complementary expertise relevant to the project includes health economics, applied microeconomics, financial econometrics; regulation and technology; privacy, fundamental rights and Internet, and cybersecurity. All project members have, in their own field, outstanding track records and most have experience with multidisciplinary research. Besides this project, the collaboration between TU/e and TiU is facilitated by their joint data science initiative, banking and insurance being defined as a strategic area of interest. On the academic leadership side of the project:

- *Mykola Pechenizkiy* is chairing the Data and Artificial Intelligence cluster. He is leading trustworthy and responsible AI research program.
- *Bas Werker* has extensive academic leadership experience and currently works on various projects on the interplay of statistics, financial econometrics and finance and insurance. In particular, he's involved in Netspar and the committee to provide longevity predictions for the Dutch Actuarial Society.
- *Jack van Wijk* was leading the visualization group at TU/e until 2022. His group has done award-winning research on visualization. He has tight collaboration with industry through joint projects, for instance with Philips, Thales, and with spin-off companies MagnaView and SynerScope (not involved in the RATE-Analytics project). He was scientific director of the Data Science Center Eindhoven (DSC/e) from 2018 to 2020.

To represent the application domain, RATE-Analytics involved two large financial institutions in The Netherlands to provide real-world, societally relevant use cases. This ensures the research within the project generalizes and addresses a real need in industry.

- *Achmea* is a leading insurance company in The Netherlands, providing various insurance products, including life, health, and property and casualty insurance. Since the company provides a significant social value, it is under heavy regulatory scrutiny. They have a keen interest in solutions to help built trust in machine learning models.

- *Rabobank* is a Dutch bank that provides a range of financial services to individuals and businesses. It is one of the largest banks in the Netherlands. Rabobank has a global presence and operates in several countries around the world, offering a variety of banking services including loans, savings accounts, and investment products. They are specifically interested in fraud detection and predictive analytics for Know Your Customer.

In all steps along the process, Achmea and Rabobank were closely involved. The choice of specific cases and problems to address was often based on discussions with experts at these companies. For most projects, real-world data was provided to make sure our solutions generalized to their application domain, and the solutions were tested in practice at these companies. This provided actionable insights for the companies and strengthened the motivation for the academic results.

It has become evident that meaningful application of machine learning-based solutions across a wide range of considered and envisioned use cases requires certain explainability for different stakeholders, including domain experts, data scientists, and customers. We zoom in into visual analytics solutions we developed throughout the project that facilitated it and and demonstrated further promise in valorization of research results.

4 Key results

To bring the human expert in the loop and address the lack of reliability and trustworthiness of current approaches, we explored interactive visualization for machine learning. Each of our works is motivated by concrete questions from data scientists at the industrial partners, and each system has also been tested and applied in practice at projects within the company. We learned that there is no silver bullet to solve all problems. Different perspectives require different solutions, ranging from local explanation of single predictions to global explanations of the entire model. Figure 2 provides an overview of how each perspective is addressed.

4.1 Technical results

We first designed and developed ExplainExplore [8]: an interactive explanation system to explore explanations of individual predictions (i.e., local). For each explanation, it provides context by presenting similar predictions, and showing the impact of small input perturbations. We recognize many different explanations may exist that are all equally valid and useful using traditional evaluation methods. Hence, we leveraged the domain knowledge of the data scientist to determine which of these fit their preference. The local perspective is particularly useful for **decision support** and **justification** applications. In a use case with data scientists from the debtor management department at Achmea, we showed the participants could effectively use explanations to diagnose a model and find problems, identify areas where the model can be improved, and support their everyday decision-making process. To ensure these contributions can be broadly applied, we also built a software library [6] that enables interoperability with a wide range of different languages, toolkits, and enterprise software.

Next, we proposed the Contribution-Value plot [9, 10] as a new elementary building block for interpretability visualization, showing how feature contribution changes for different feature values. It provides a perspective in between local and global, as the model behavior is shown for all instances, but visualized on a per-feature basis. This perspective primarily helps to **diagnose** unexpected model behavior. In a quantitative online survey with 22 machine learning professionals and visualization experts, we show our visualization increases

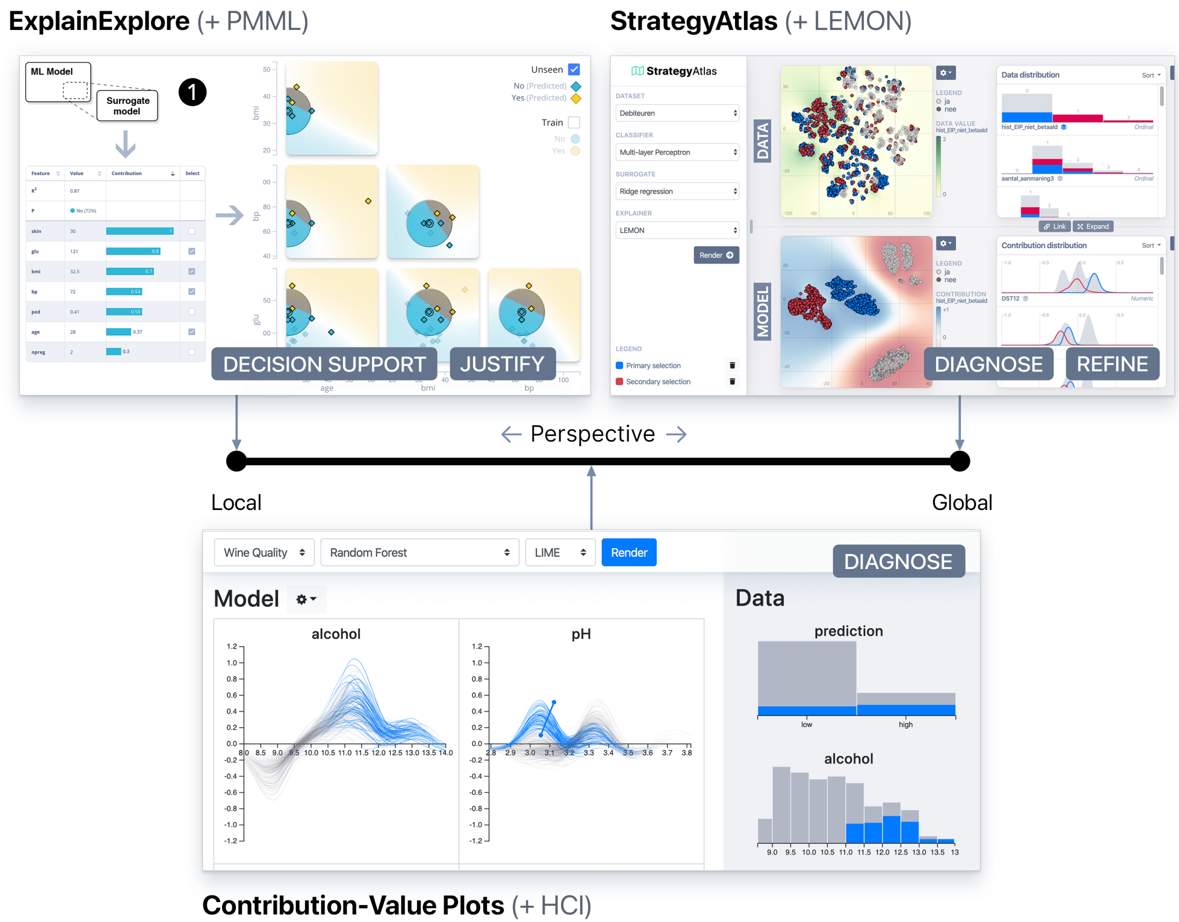


Figure 2 Overview of the different perspectives and corresponding systems developed in the RATE-Analytics project. ExplainExplore (←) provides local explanations of individual predictions with a bar chart, and provides context with class activation heatmaps; CV plots (↑) explain how feature improtance changes for different feature values; and StrategyAtlas (→) provides a global explanation of the entire model through model strategies that highlight different model behavior for subgroups. The three core VA solutions are accompanied by complementary contributions outside the visualization domain.

correctness and confidence and reduces the time needed to obtain an insight compared to previous techniques. This work highlighted that a small difference in feature importance techniques can result in a large difference in interpretation, and warranted a follow-up human-computer interaction contribution to characterize the data scientists' mental model of explanations [12], and explore the differences between existing techniques.

Finally, we designed and developed StrategyAtlas [11]: a visual analytics approach to enable a global understanding of complex machine learning models through the identification and interpretation of different model strategies. These model strategies are identified in our projection-based StrategyMap visualization. Data scientists are enabled to ascertain the validity of these strategies through analyzing feature values and contributions using heat maps, density plots, and decision tree abstractions. This global perspective is especially beneficial to **diagnose** and **refine** models. In collaboration with Achmea, we applied the system in a real-world project for automatic insurance acceptance. This showed that professional data scientists were able to understand a complex model and improve the production model based on these insights. As computing the local feature importance values for an entire dataset is computationally expensive, we complemented this work with an algorithmic contribution called LEMON [7] to improve the faithfulness of explanation results, which enables us to significantly speed up computations of StrategyMap projections.

Although the individual visualizations introduced our work are simple in nature, we have shown that the visualizations *combined* form a strong visual encoding that enables data scientists to gain an understanding of machine learning models which was not possible before. This enables data scientists to diagnose, refine models, and support decision-making and justification of predictions. Its not the individual visualizations, but the full picture - the entire interactive workflow - that yields insights.

In addition to our interactive visualization contributions, we also studied how to deal with data that is not independent and identically distributed, i.e. so-called non-IIDness. When applying predictive analytics in real-world applications, complex behavior exhibits lots of heterogeneity, including temporal dynamics and coupling relationships.

Effective anomaly detection in financial transaction networks is particularly challenging because there exist collective fraudulent behaviors at the level of subgraphs rather than individual nodes. A ring structure for money laundering and a tree structure for pyramid schemes would be common examples. In practice it is difficult to decide which features are more representative beforehand. We introduced a subgraph anomaly detection framework that allows to preserve both the local structure of subgraphs and the global structure of entire network by making use of global roles and local connections of nodes [22]. We demonstrated on both synthetic and real-world financial transaction network datasets the effectiveness of learning subgraph embeddings without requiring any prior knowledge and detecting anomalous subgraphs. In [21] we proposed a novel way of employing attention-based deep residual modeling approach that can effectively detecting anomalous nodes in attributed networks [18] by capturing the sparsity and nonlinearity, reducing the adverse effect from anomalous nodes and preventing so-called over-smoothing in representation learning. Our experiments on several real-world attributed networks demonstrated the effectiveness of detecting anomalies in these settings.

Learning from multiple time-series over an unbounded time-frame requires machine learning techniques to continuously learn from data that evolves over time, exhibits concept drifts. We explored ensembles learning frameworks that allow to capture changes in the concepts [19]. One of challenges is that supervised learning for modeling such second order dynamics is often not fully applicable since the true labels (e.g. default clients, fraudulent behavior etc) become known with considerable delay. We studied the potential of employing labelless concept drift detection, visual inspection, and explanation with use local feature importance that is used for constructing a new representation allowing effective drift detection while maintaining acceptable level of false positives [25].

All of the described work has the potential to be widely applied in various industries and domains where machine learning is used. This includes other projects at Achmea and Rabobank, but also other domains beyond the scope of RATE-Analytics. Especially in high-stakes decision-making, where models make decisions that have a critical impact on real people, it is not sufficient to base decisions on the prediction score of the model alone. To qualitatively ascertain whether desiderata such as fairness, privacy, and trust are met, our results can help to verify the behavior of models [13], through enabling understanding of complex models. To reach these other domains, we made a promotional website[1] containing simple explanations of our work, along with conference presentations, online demos and source code.

4.2 Collaboration results

The RATE-Analytics project has a rather unique positioning: aiming to advance the research area of predictive analytics, and to have an immediate impact on its acceptance in the industry. Apart from our academic contributions, the tight collaboration between universities and industry proved most fruitful.

The researchers gave several presentations at Achmea and Rabobank. This includes data science teams, senior management, innovation managers, and the Chief Information Officer (CIO) of Achmea. These presentations contributed to the ongoing internal discussion on how to use AI responsibly within Achmea and across all insurance companies in the Netherlands (through the "Verbond Van Verzekeraars").

It also enabled the industry partners to immediately incorporate and apply state-of-the-art research. The debtor management department at Achmea used ExplainExplore to analyze their model built to predict the effectiveness of a debtor contact strategy, and found that the model used several attributes in the data they did not expect were relevant for the prediction. This insight could helpt them refine their current model. Next, data scientists at Achmea used StrategyAtlas with an automatic acceptance model for car insurances. They system helped them realize their model did not detect bad actors and removed them, but rather looked at the characteristics of good customers, which was contrary to their prior beliefs.

Conversely, Achmea and Rabobank provided relevant use cases for evaluating our approaches, such as the ones mentioned above. This provided the academic partners with a strong motivation for the projects and evaluation of the published work.

[1] `https://explaining.ml`. When offline, an archived version of the website can be found at `https://web.archive.org/web/20230515141514/https://explaining.ml/`.

5 Future prospects

Since the RATE-Analytics project aimed for developing generic technology, there is a potential of utilizing the outcome of the project in other domains beyond banking and insurance. E.g. the three core research problems we formulate are relevant to the area of big data in medical diagnostics and predictive maintenance tasks in production industries.

In general, the results of this project can help companies to develop more reliable (high detection rate), more cost-effective (helping human analysts with the right tools to deal with typically large number of false alarms), and more transparent and interpretable analytics.

To explore this idea further, we applied and were granted admission to the NWO Take-off programme (phase 1). Participation in this programme and the corresponding €40.000 subsidy will help us investigate the feasibility of a startup company to help businesses understand and explain complex machine learning models. We will mainly focus on conducting interviews with potential customers and end-users, to ascertain 1) if black box models are indeed considered a problem, and by who; 2) which roles in the organisation are responsible for the buying decision (decision-making unit); 3) if our solution sufficiently addresses the customers problem (i.e., if our explanations can yield business value in practice), and 4) to find commercial commitment from an initial launching customer.

The envisioned startup helps companies that conduct data science for high-stakes decision making who want to understand machine learning models by 1) reducing the risk of (invisible) unethical decisions, and 2) reducing the risk of non-compliance with current (GDPR) and upcoming (AI Act) regulations that will restrict how companies can apply data science. This startup enables us to continue developing our solutions in industry context, make available our solutions to a broader audience, thus have a bigger impact on society, and offer Achmea and Rabobank continued support and benefit from the RATE-Analytics project results.

The RATE-Analytics project provided several benefits to these valorization efforts. The fact that the RATE-Analytics project addresses real problems in industry provides some immediate problem validation, a key first step in the development of a startup company [3]. Next, because the research contributions were all evaluated in practice at the companies, we could immediately assess the Technology Readiness Level (TRL) at grade 5. This is a helpful method for estimating the maturity of technologies, and part of the NWO Take-off application process.

To investigate the practical and commercial viability of our business model, we plan to first identify potential customers. Next, through initial meetings with these customers we plan to get early feedback on the current prototypes and business idea. While the current prototypes have been validated in the expected customer segment at Achmea, we need to make sure these prototypes are suitable for other customers too. We anticipate different customers need different levels of detail, and the prototype needs tailoring for those specific needs. Therefore, as a main work within this project, we plan to prepare a convincing and understandable demonstration of our platform to be used as the basis for the future startup.

We are looking forward to developing and expanding our ideas and technology further. The RATE-Analytics project, aimed at developing new technology in close cooperation with companies from finance, was a crucial first step. We hope that next steps will lead to much impact across a variety of application domains.

References

1 Mihael Ankerst, Martin Ester, and Hans-Peter Kriegel. Towards an effective cooperation of the user and the computer for classification. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 179–188. ACM, 2000. doi:10.1145/347090.347124.

2 Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Zelezný, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III*, volume 8190 of *Lecture Notes in Computer Science*, pages 387–402. Springer, Springer, 2013. doi:10.1007/978-3-642-40994-3_25.

3 Steve Blank. *The four steps to the epiphany: successful strategies for products that win.* John Wiley & Sons, 2020.

4 Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, 2018. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

5 Declan Butler. When Google got flu wrong. *Nat.*, 494(7436):155–156, 2013. doi:10.1038/494155a.

6 Dennis Collaris. sklearn-pmml-model: Machine learning portability and interoperability using PMML, 2022. Library available at https://github.com/iamDecode/sklearn-pmml-model.

7 Dennis Collaris, Pratik Gajane, Joost Jorritsma, Jarke J van Wijk, and Mykola Pechenizkiy. LEMON: Alternative sampling for more faithful explanation through local surrogate models. In *Advances in Intelligent Data Analysis XXI: 21st International Symposium on Intelligent Data Analysis (IDA 2023)*, pages 77–90. Springer, 2023. doi:10.1007/978-3-031-30047-9_7.

8 Dennis Collaris and Jarke J van Wijk. ExplainExplore: Visual exploration of machine learning explanations. In *Proceedings of the 2020 IEEE Pacific Visualization Symposium (PacificVis)*, pages 26–35. IEEE, 2020. doi:10.1109/PacificVis48177.2020.7090.

9 Dennis Collaris and Jarke J. van Wijk. Machine learning interpretability through contribution-value plots. In Michael Burch, Michel A. Westenberg, Quang Vinh Nguyen, and Ying Zhao, editors, *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*, pages 4:1–4:5. ACM, 2020. doi:10.1145/3430036.3430067.

10 Dennis Collaris and Jarke J. van Wijk. Comparative evaluation of contribution-value plots for machine learning understanding. *Journal of Visualization*, 25(1):47–57, 2022. doi:10.1007/s12650-021-00776-w.

11 Dennis Collaris and Jarke J. Van Wijk. StrategyAtlas: Strategy analysis for machine learning interpretability. *IEEE Transactions on Visualization and Computer Graphics*, 29(6):2996–3008, 2023. doi:10.1109/TVCG.2022.3146806.

12 Dennis Collaris, Hilde J. P. Weerts, Daphne Miedema, Jarke J. van Wijk, and Mykola Pechenizkiy. Characterizing data scientists' mental models of local feature importance. In *NordiCHI '22: Nordic Human-Computer Interaction Conference, Aarhus, Denmark, October 8 - 12, 2022*, pages 9:1–9:12. ACM, 2022. doi:10.1145/3546155.3546670.

13 Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

14 Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017. doi:10.1609/aimag.v38i3.2741.

15 Philipp Hacker and Jan-Hendrik Passoth. Varieties of AI explanations under the law. from the GDPR to the aia, and beyond. In Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek, editors, *xxAI - Beyond Explainable AI - International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria,*

Revised and Extended Papers, volume 13200 of *Lecture Notes in Computer Science*, pages 343–373, Cham, 2020. Springer. doi:10.1007/978-3-031-04083-2_17.

16 Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven Mark Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In Stephen A. Brewster, Geraldine Fitzpatrick, Anna L. Cox, and Vassilis Kostakos, editors, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, page 579. ACM, 2019. doi:10.1145/3290605.3300809.

17 Tianjin Huang, Vlado Menkovski, Yulong Pei, Yuhao Wang, and Mykola Pechenizkiy. Direction-aggregated attack for transferable adversarial examples. *ACM J. Emerg. Technol. Comput. Syst.*, 18(3):60:1–60:22, 2022. doi:10.1145/3501769.

18 Tianjin Huang, Yulong Pei, Vlado Menkovski, and Mykola Pechenizkiy. Hop-count based self-supervised anomaly detection on attributed networks. In Massih-Reza Amini, Stéphane Canu, Asja Fischer, Tias Guns, Petra Kralj Novak, and Grigorios Tsoumakas, editors, *Proceedings of European Conference on Machine Learning and Knowledge Discovery in Databases - , ECML PKDD 2022, Part I*, volume 13713 of *Lecture Notes in Computer Science*, pages 225–241. Springer, 2022. doi:10.1007/978-3-031-26387-3_14.

19 Samaneh Khoshrou and Mykola Pechenizkiy. Adaptive long-term ensemble learning from multiple high-dimensional time-series. In *Discovery Science: 22nd International Conference, DS 2019, Split, Croatia, October 28–30, 2019, Proceedings 22*, pages 511–521. Springer, 2019. doi:10.1007/978-3-030-33778-0_38.

20 Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 427–436. IEEE Computer Society, 2015. doi:10.1109/CVPR.2015.7298640.

21 Yulong Pei, Tianjin Huang, Werner van Ipenburg, and Mykola Pechenizkiy. ResGCN: attention-based deep residual modeling for anomaly detection on attributed networks. *Machine Learning*, 111(2):519–541, 2022. doi:10.1007/s10994-021-06044-0.

22 Yulong Pei, Fang Lyu, Werner van Ipenburg, and Mykola Pechenizkiy. Subgraph anomaly detection in financial transaction networks. In Tucker Balch, editor, *ICAIF '20: The First ACM International Conference on AI in Finance, New York, NY, USA, October 15-16, 2020*, ICAIF '20, pages 18:1–18:8, New York, NY, USA, 2020. ACM. doi:10.1145/3383455.3422548.

23 Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144. ACM, ACM, 2016. doi:10.1145/2939672.2939778.

24 Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. Interacting meaningfully with machine learning systems: Three experiments. *International Journal of Human-Computer Studies*, 67(8):639–662, 2009. doi:10.1016/j.ijhcs.2009.03.004.

25 Shihao Zheng, Simon P. van der Zon, Mykola Pechenizkiy, Cassio P. de Campos, Werner van Ipenburg, and Hennie de Harder. Labelless concept drift detection and explanation. In *Proceedings of NeurIPS 2019 Workshop on Robust AI in Financial Service*, 2019.

26 Indrė Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. In *Big Data Analysis: New Algorithms for a New Society*, pages 91–114. Springer, 2016.

Digital Art Technical Sources for the Netherlands: Integration and Improvement of Sources on Glass for a Sustainable Future – Art DATIS

Carlotta Capurro[1]
Utrecht University, The Netherlands

Vera Provatorova
University of Amsterdam, The Netherlands

Marieke Hendriksen
Huygens Institute, Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands

Evangelos Kanoulas
University of Amsterdam, The Netherlands

Sven Dupré
Utrecht University, The Netherlands

Abstract

Art DATIS (Digital Art Technical sources for the Netherlands: Integration and improvement of sources on glass for a Sustainable future) is a five-year research project (2018-2023) within the Netherlands Organisation for Scientific Research's (NWO) Big Data / Digital Humanities program. The project is a collaboration between the Universities of Utrecht and Amsterdam, RKD Netherlands Institute for Art History, the Vrij Glas Foundation, and Picturae. The project investigates how to approach the automatic transcription and documentation of heterogeneous archival resources. The central object of the project is the archive of the Dutch glass artist Sybren Valkema (1916–96). Documents were digitised, and their content was made searchable through the processes of OCR and HTR. Through digitisation and the analysis of archival documents, the project aims to understand how traditional knowledge and practices of glassmaking were innovated during the twentieth century.

2012 ACM Subject Classification Applied computing → Fine arts; Computing methodologies → Neural networks

Keywords and phrases Digital Humanities, Archives, Digitisation, Datafication, Digital Art History, Technical Art History

Digital Object Identifier 10.4230/OASIcs.Commit2Data.2024.9

Funding NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), grant number NWO Smart Culture – Big Data/Digital Humanities, 314-99-301.

1 Introduction

Digital Art Technical sources for the Netherlands: Integration and improvement of sources on glass for a Sustainable future (Art DATIS) is a five-year research project launched in 2018 and funded by the Netherlands Organization for Scientific Research (NWO) within the Big Data / Digital Humanities program [8]. With this grant, NWO aimed at funding proposals that contribute to the development of research in the field of creative industries and Big Data while boosting the research in the field of Digital Humanities [18]. The Art DATIS project sees the collaboration of five partners: the Universities of Utrecht (UU) and Amsterdam

[1] Corresponding Author

Commit2Data.
Editors: Boudewijn R. Haverkort, Aldert de Jongste, Pieter van Kuilenburg, and Ruben D. Vromans; Article No. 9; pp. 9:1–9:11
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

(UvA), the Netherlands Institute for Art History (RKD), the Vrij Glas Foundation, and the Dutch heritage digitisation company Picturae. It aims to understand how traditional knowledge and practices of glassmaking were innovated during the twentieth century by analysing the work of one of the key actors in this process, the Dutch glass artist Sybren Valkema (1916–96).

Sybren Valkema was one of the representatives of the Studio Glass movement that, in the Netherlands, took the name of Vrij Glas. Spreading internationally from the US during the 1960s, Studio Glass promoted the transformation of traditional glassmaking practice, encouraging the use of glass as an artistic medium. The movement also had a significant impact on the circulation of knowledge about glass and glassmaking techniques. Believing in knowledge sharing and collaborative improvement, many artists adhering to the Studio Glass movement took up educative roles in art academies and universities, establishing courses to teach glass-blowing as a form of art [10] [9]. In line with the Studio Glass philosophy, Valkema combined glass-blowing with an intense career as a teacher, working at the Dutch Institute for Art and Crafts (Instituut voor Kunstnijverheidsonderwijs), later renamed Gerrit Rietveld Academy.

Central to Art DATIS research is Valkema's personal archive. Sources containing historical recipes and techniques for glass production, as well as images and object documentation, are currently digitised by museums and research institutions around the world. The Valkema's archive contains over 103.000 pages documenting the private and professional life of the artist, including teaching materials, letters, designs, sketches, descriptions of technical processes and many glass recipes, which have been tested and revised with handwritten notes [Figure 1]. Due to the importance of these resources for the study of glass art in the twentieth century, Metamorfoze, the Netherlands' national programme for the preservation of paper heritage, financed the conservation and digital preservation of the archive. Between 2013 and 2015, the archive was digitised thanks to the joint efforts of RKD and the Rakow Research Library of the Corning Museum of Glass [23].

Through digitising and enriching the documents in Valkema's archive, the Art DATIS project worked towards creating a comprehensive digital resource for the study of contemporary glassmaking, allowing scholars, artists and practitioners to investigate the use of glass as an artistic material through time. One of the main issues it had to face was implementing a robust process for the datafication of archival resources. The concept of datafication is not limited to its literal meaning of transforming something into data but includes the social, cultural and political implications that such a process has on our lives [17]. When dealing with archival resources, their transformation into data requires attentive consideration of the possible biases and the implications they might have on our understanding of the past [13].

As a project developed in the Digital Humanities, Art DATIS has benefitted from two separate yet interlocking disciplinary approaches: information retrieval and digital art history. First, the project investigated the best approach to make the digital documents preserved in Valkema's archive accessible and usable as digital resources. Images were converted into machine-encoded text with the support of Text Recognition software (handwritten pages were processed using Trankribus, while typed documents were processed with the OCR engine Tesseract), and then the content was extended, enriched, and linked with open data. Second, this material was used to investigate Valkema's international network and shed light on the role of these relationships in his artistic production and the spread of Studio Glass in Europe.

This contribution aims to illustrate the developments within the ArtDATIS project and is structured as follows. The second section offers a historical introduction to the Studio Glass movement and the work of Sybren Valkema. The documents digitised and analysed

Figure 1 Example of the documents preserved in Sybren Valkema's archive. They include drawings, school notes, working notebooks, letters and booklets.

within the ArtDATIS framework are contextualised historically to illustrate better their relevance for the study of the international development of the Studio Glass movement. The third section describes the development of the project plan and the partners' role within the project's consortium, focussing on each member's role in the success of the project. The fourth and fifth sections describe, respectively, the issues faced and the solutions developed within the ArtDATIS project. Finally, the last section offers some conclusive remarks and lessons learned.

2 Art Historical Background: Sybren Valkema and the Vrij Glass Movement

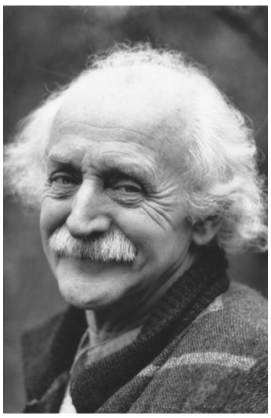


Figure 2 Sybren Valkema (Photo credit: Anna Carlgren, 1985).

From an art historical perspective, the Art DATIS project moved from the standpoint that the production of free glass saw significant technological innovations in the twentieth century. Due to the high temperatures and complex infrastructures that were necessary to melt glass, most of the early-twenty-century glass production happened within industrial facilities, where architects and designers were often invited to develop innovative and more sophisticated shapes capable of attracting the public's tastes [19]. With the intent of emancipating glass artists from industrial facilities in the early 1960s, Harvey Littleton (1922-2013) and Dominick Labino (1910-1987) prototyped an innovative studio furnace. It allowed artists to experiment freely with glass materials to produce aesthetically pleasant pieces [2].

In 1962, upon the invitation of the museum's director Otto Wittmann (1911-2001), Littleton organised two workshops on glass-blowing at the Toledo Museum of Glass. American artists and ceramists were invited to attend the workshops, experimenting with the studio furnace for melting and blowing glass. These events were considered a success for the knowledge they produced in setting up and running a studio production of glass, and are formally recognised as the inaugural events of the Studio Glass movement[16]. When, in 1964, the studio furnace was shown at the First World Craft Congress held in New York at Columbia University, Littleton's ideas encountered the support of a large number of artists and makers, determining the worldwide spread of the Studio Glass movement.

The Studio Glass movement encouraged artists to work independently in private studios, fostering a sense of artistic freedom and exploration. Within the glass-blowing studios, artists engaged in extensive research to push the boundaries of the glass medium, experimenting with new techniques and forms that emphasised individual expression and creativity. At the same time, the movement embraced education as a lever for transforming the art world, with universities playing a pivotal role in disseminating technical knowledge on glassmaking and cultivating a new generation of glass artists. In 1964, Littleton set up the first course on studio glass at the University of Wisconsin-Madison to promote glass-blowing among students. Between the second half of the 1960s and the early 1970s, the number of American universities offering courses in glass blowing grew considerably [16].

Sybren Valkema was among the participants at the 1964 World Craft Congress [Figure 2]. Despite his long and close collaboration as a design teacher with the Royal Leerdam Crystal, a Dutch factory producing glassware, Valkema had no direct experience in blowing glass [24]. In New York, he was fascinated by Littleton's demonstration and could experiment for the first time with the blowpipe. Once back in the Netherlands, he set up a glass-blowing laboratory on the premises of the Rietveld Academy in Amsterdam, building the first studio furnace in the Netherlands. In 1969, he inaugurated the first free glass curriculum, open to students of different disciplines [Figure 3]. During his life, Valkema experimented with glass and glass-blowing, studying historical sources and innovating them with newly developed materials and techniques, crucially contributing to the innovation the Studio Glass movement brought up.

Through his artistic research, Valkema established long-lasting friendships with most of the artists involved in the Studio Glass movement. Located in several countries worldwide, from the US to Australia and Japan, passing from the United Kingdom, Germany, and the Czech Republic, Valkema entertained a close epistolary exchange with artists such as Harvey Littleton, Erwin Eisch (1927-2022), Sam Herman (1936-2020), Stanislav Libenský (1921-2002) and Jaroslava Brychtová (1924-2020). With them, he exchanged information on techniques, designs, and events from the glass-blowing community. At the same time, students from the various glass courses were invited to spend a period abroad practising in the workshop of a different academy, generating a dynamic exchange of knowledge and ideas. The documents

Figure 3 Sybren Valkema (in the background) observes his son Durk blowing glass, assisted by Anna Carlgren, 1982 (Archive Sybren Valkema).

preserved in the Valkema archive, among which many witness the correspondence the artist had with friends, colleagues, and students, are invaluable for studying this lively environment. Therefore, one of the main research lines developed within the Art DATIS project investigated how knowledge circulated within the Studio Glass movement by reconstructing the network of Valkema's contacts.

3 The Consortium

In its aims, Art DATIS aligned with Studio Glass' ambitions of making glass-related knowledge widely available, this time using the digital medium. This was mainly made possible by the specific composition of the consortium. It was formed by five partners: two from academia (Utrecht University and the University of Amsterdam) and three public (RKD) and private institutions (Vrij Glas Foundation and Picturae). Within the consortium, each member had specific competencies in art history (UU, RKD and Vrij Glass Foundation), data science (UvA), data management (UvA, RKD and Picturae) and digital humanities (UU).

The project's idea stemmed from the collaboration between Utrecht University and RKD within the framework of the ARTECHNE project [1]. ARTECHNE: Technique in the Arts 1500–1950 was a European Research Council-funded project hosted between 2015 and 2021 at Utrecht University and the University of Amsterdam. The project developed an online database of art technological sources containing very heterogeneous data: fully searchable texts in six languages from the period 1500 to 1900, images, records on historical people, geotags, and historical and current names of artistic materials and techniques [14]. Intending to provide for the long-term sustainability of the database, ARTECHNE partnered with

RKD, which added it to the list of digital resources offered within its search engine RKD Explore. This collaboration led RKD, which in the meantime had acquired Sybren Valkema's digital archive, to propose linking the historical sources in its collection with the documents in the archive and other existing big data in the field of art history and technical art history. This idea would have contributed to enriching Valkema's digital archive, falling within the scope of the call on big data and smart culture issued by the NWO. Thanks to Art DATIS, the ARTECHNE database was enriched with new entries on glass and glassmaking (i.e. historical recipes and descriptions of glassmaking procedures from historical sources).

Art DATIS, therefore, was the natural follow-up of the digitisation of Valkema's archive. In the process of digitisation of the archival resources, the work performed by the company Picturae, which specialised in collaborating with cultural heritage institutions to produce high-quality digital duplicates, was crucial. Due to its competence with the management of digital cultural data and its knowledge of the archive, Picturae was involved in the development of the Art DATIS project.

With the aim of forming a robust cohort of partners, RKD contacted the Vrij Glas Foundation, the legal owner of the Valkema archive, and involved them in the project. The Foundation is directed by Durk Valkema, Sybren's son, and his wife, Anna Carlgren. They both shared an artistic career working with glass, learning glassblowing at the Rietveld Academy under Sybren's supervision, and witnessing the evolution of the Studio Glass Movement in the Netherlands and internationally. As the gatekeeper of Valkema's heritage, the Foundation was enthusiastic about the prospect of making the archive widely available, emphasising the continuity between its digitisation and the Studio Glass support of open and free circulation of technical knowledge about glass. The intimate knowledge of the archival material and the first-hand experience of Valkema's glass-blowing practice made the presence of Vrij Glass in the consortium an invaluable resource.

Within the project, Utrecht University and the University of Amsterdam supervised the scientific research development, collaborating to make the Valkema archive digitally accessible and searchable. The research team was composed of three researchers and three supervisors. The project trained two PhDs, one in the field of art history and one in data science. They were based, respectively, at Utrecht University (History and Art History Department) and the University of Amsterdam (Informatics Institute). Additionally, a Postdoctoral researcher (Utrecht University, History and Art History Department) joined the project to facilitate the process of datafication of the archival resources. The research team closely collaborated during the whole development of the project, benefitting from the support of the other members of the consortium. Each partner offered its competencies on specific aspects of the work. The Vrij Glas Foundation, which inherited Valkema's artistic practice and is in charge of preserving his archive, offered support with its know-how on glassmaking; RKD offered the digital space to protect and enrich digital resources, granting their long-term sustainability, while the digital cultural heritage company Picturae provided advice on technical solutions and best practices in digital preservation.

4 Problem addressed

Art DATIS represents an outstanding case study for approaching the mass digitisation and automatic documentation of heterogeneous archival resources. The project investigated how to efficiently deal with the datafication of heterogeneous archival resources, connecting them with existing digital art technical sources on artistic glass production, such as object documentation, technical texts, images, and research data. Therefore, the project's ultimate goal was to contribute to the creation of a comprehensive art historical resource about glass and glassmaking, allowing scholars, artists and practitioners to investigate the use of glass as an artistic material through time.

The diverse nature of the resources in the archive requires each a specific treatment to be transformed from simple digital images into meaningful data resources. The first challenge Art DATIS had to overcome was finding a way to describe and document each entry in the collection to process it adequately. This includes creating a set of metadata containing enough information to allow retrieving the resource in the appropriate context. As a matter of fact, when the archive was digitally remediated, resources were grouped into broad thematic categories, which offered those who approached the collection a general idea of their context but needed to be renamed and documented in detail.

The lack of prior information about the resources also represented an optimal opportunity to experiment with innovative archival practices and develop a system allowing resource documentation with minimal content interpretation. Scholars have closely scrutinised the practices of object documentation applied by archives and cultural heritage institutions, recognising that, at times, they have been subjected to cultural and political biases, which affect the production and dissemination of new knowledge [12] [3] [4]. For this reason, Art DATIS worked to implement a rich documentation system based on the very content of each document, aiming to reduce their interpretation.

To address these issues, the project had to investigate the technical aspects of datafication. The archive comprises various heterogeneous resources, including handwritten and typed documents, designs, and pictures. Art DATIS focussed on written resources, aiming at making them searchable. To make it possible, documents had to be categorised and analysed. It was first necessary to separate visual documentation from written resources; then, the latter had to be processed with the most appropriate tool to extract the textual information.

To make content searchable, we had to convert images into machine-encoded text, a process named Text Recognition [11]. As the following section will better describe, in practice, this required developing an algorithm capable of automatically identifying written sources and separating them from visual contents. Then, a second algorithm was designed to distinguish typed from handwritten documents. The former type of documents were then transcribed using an OCR tool, while the latter required HTR software. Once transcribed, the documents became searchable, and their content could eventually be used to generate automatic metadata. To facilitate this process, an algorithm allowing automatic entity recognition and linking was developed [22].

The joint collaboration of art historians and data scientists was crucial during the whole development of the project. It allowed for research in the field of digital humanities, developing a tool at the service of the community of scholars and artists. The archival resources, enriched with data generated by external resources and made available through the RKD servers, will allow researchers and practitioners to extensively investigate the transformation of glass production throughout time, browsing old and new recipes to discover the role of traditional practice in contemporary glass-blowing.

5 Key results

To address the issues at the core of the research project, it was first necessary to develop a strategy to sort out resources according to their type (i.e. drawings, pictures, typed texts or handwritten texts) and process each of them with the most appropriate tool. The focus of Art DATIS was primarily documentary resources. Therefore, we mostly worked to develop

Table 1 Evaluation of OCR's results.

type	examples	number of pages in the sample
true positive	typed pages; some mixed pages (where the majority of the text is typed but some handwritten notes or drawings are present)	27 (24 typed, 3 mixed)
false positive	handwritten pages with a very clear writing style (not a problem, can be OCRed); some mixed pages (where the amount of handwritten text is sufficient to make HTR necessary)	5 (1 handwritten, 3 mixed, 1 back)
true negative	handwritten pages, drawings, backsides, blank pages, some mixed pages (where only a small amount of typed text is present)	60 (14 handwritten, 35 backsides, 5 blank pages, 6 mixed)
false negative	typed pages missed by the classifier	8 (7 typed, 1 mixed)

optimal OCR and HTR solutions. To identify which documents are typed pages that can be processed with OCR algorithms, we developed a rule-based binary classifier using the following heuristic: if running the Tesseract OCR engine on a page detects Latin script, then the page is likely typed [15]. To validate this approach, we randomly sampled 100 pages from the archive and performed a manual evaluation. Out of the 100 pages, 35 were typed. The classifier was found to have 84% precision, i.e. it predicted 32 pages as typed, out of which 27 were actually typed, and 77% recall, i.e. out of all 35 typed pages, it successfully identified 27. Detailed results of the evaluation can be seen in Table 1.

On the other hand, the handwritten material needed to be processed with appropriate HTR algorithms. Art DATIS consortium used Transkribus, a platform enabling AI-powered text recognition. Transkribus relies on the principles of machine learning and employs artificial neural networks to assist researchers in training models that can read and transcribe their collections automatically. [6]. Performing HTR on the Valkema archive was particularly complex due to two main issues. First, the archive includes not only documents written by Valkema at any stage of his life, therefore presenting a variation in his handwriting, but also documents written by other people in several languages. Current HTR methods require extensive manual annotation efforts to create training data: every language and every handwriting have unique features and ideally need the use of a separate model. Transkribus suggests utilizing a minimum of 10,000 words per unique writer to effectively train a model [7]. Second, due to the fact that digital resources are undocumented, the lack of metadata about the language and author makes it impossible to automatically sort out the material for processing it with a specific model.

Using Transkribus, it was possible to train different language models and experiment with their efficiency in transcribing the collection. Due to the lack of metadata on the language used in the documents, we identified the more common languages used in the archive (Dutch, English, and German) and developed a model capable of producing automatic transcription with an acceptable margin of error for all the languages. Therefore, we conceived a series of experiments to assess the efficiency of a multilingual, multi-author model trained on a sample of documents from the Valkema archive. Its efficiency was tested on a sample of documents randomly selected in the archive and compared to the results produced by models trained on documents in a single language. Table 2 shows the monolingual and multilingual models we have trained and their performances. The multilingual model performed adequately in comparison with the monolingual ones, allowing for the automatic transcription of the whole archive [5].

Once all the images were converted into textual information using OCR and HTR, the archive became searchable. The second main challenge faced within Art DATIS was then designing a system to automatically identify meaningful entities embedded in the text and link them with existing open data. Identifying named entities, such as people and cities, is a useful tool for digital archives: it allows for the improvement of the metadata and

Table 2 Overview of the HTR models trained during the experiment and their performances.

Language	Training details	Title & Transkribus ID	Performance during training (average CER)	Performance of the model calculated on our collection (average CER)
EN	base model: Transkribus English Handwriting M3 (ID 37646) training data: 87 pages (10% used for validation)	ArtDATIS_English_latest (53449)	CER on train: 4.7% CER on validation: 2.1%	CER on test: 11.1% (STD 12.8%)
NL	base model: Transkribus Dutch Handwriting M2 (ID 45422) training data: 215 pages (10% used for validation)	ArtDATIS_Dutch_latest (53462)	CER on train: 9.8% CER on validation: 6.1%	CER on test: 18.9% (STD 9.5%)
Multi	base model: Transkribus Dutch Handwriting M2 (ID 45422 training data: EN – 87 pages NL – 215 pages (10% in each language is used for validation)	ArtDATIS_Dutch_English (53444)	CER on train: 6.6% CER on validation: 3.2%	CER on test: 20.1% all languages (STD 14.8%) 12.5% EN (STD 13.1%) 30.7% DE (STD 14.9%) 17.1% NL (STD 9.5%)

makes the documents easily searchable. Entity linking takes the process one step further by connecting unstructured text with a structured knowledge base, such as Wikipedia or Wikidata. Entity linking in digital archives comes with a number of challenges: (a) OCR and HTR errors confuse entity identification and linking algorithms, while (b) entities of the past are overshadowed by entities of the present due to popularity bias in entity identification and linking algorithms. We studied the latter problem [20] and proposed solutions for robustifying the entity linking algorithm [21].

The automatic identification of entities related to people allowed us to generate an original visualisation of the information contained in the archive: Valkema's social network [Figure 4]. Such a visualisation shed light on how innovation in glassmaking travelled across countries in the second half of the Twentieth century through the social connection of artists and the exchange of students.

6 Conclusion and looking ahead

One of the most remarkable results of the Art DATIS project is the creation of a simple procedure for the datafication of mass digitised archives, which can be easily imitated by cultural heritage institutions. Most of the archival collections are, in fact, constituted of heterogeneous material that needs to be processed with ad hoc instruments in order to become searchable. At the same time, the large number of documents that have to be processed and the amount of funding necessary for the operation make it very difficult for cultural institutions to process their entire collections. In this sense, the solutions developed within Art DATIS might represent a valuable reference for similar collections. Concretely, the multilingual HTR model trained using Transkribus will be freely available for reuse on

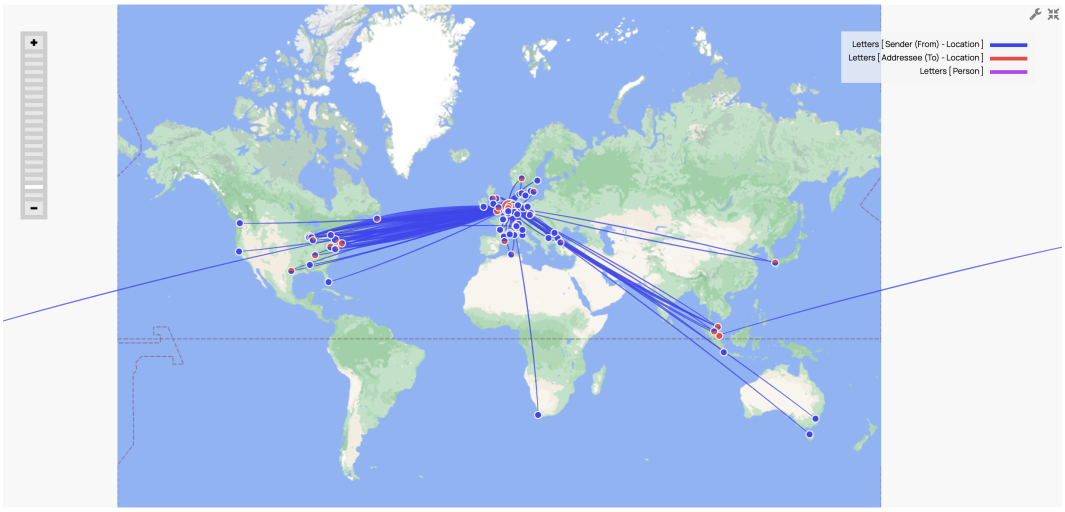


Figure 4 Visualisation of the geographical provenance of the correspondence in the Valkema's archive. Graph generated with the software Nodegoat.

the tool platform. Any institution or researcher dealing with handwritten documents will be able to use the model to transcribe their collections, saving time and resources in training a new ad-hoc model. In the same way, the entity recognition and entity linking algorithms developed to process the Valkema collection will be made available, a.o. as part of the toolset offered by the RKD online platform.

To conclude, the most important lesson that the project can offer to similar ones addressing the datafication of archival resources is to carefully estimate the time and resources necessary to reach their goals. The heterogeneous nature of archival resources makes them singular research objects, which poses specific and time-consuming issues when digitised and transformed into data. These aspects need to be taken into consideration in the project design phase in order not to jeopardise the success and financial viability of the project. In this sense, the close collaboration of experts from the humanities and technical disciplines is necessary already in the earliest phases of data creation in order to successfully complete digital humanities projects. Funding agencies should create financial opportunities for this, and as this project shows, investment in this phase of data creation can also lead to more widely useful results and tools.

References

1. ARTECHNE. ARTECHNE - Technique in the Arts, 1500-1950 | Technique in the Arts, 1500-1950: Concepts, Practices, Expertise. URL: `https://artechne.wp.hum.uu.nl`.
2. Joan Falconer Byrd and Harvey K. Littleton. *Harvey K. Littleton - a life in glass: founder of America's studio glass movement.* Skira Rizzoli, New York, 2011.
3. Fiona Cameron and Helena Robinson. Digital Knowledgescapes: Cultural, Theoretical, Practical, and Usage Issues Facing Museum Collection Databases in a Digital Epoch. In Fiona Cameron and Sarah Kenderdine, editors, *Theorizing Digital Cultural Heritage*, pages 165–191. The MIT Press, March 2007. `doi:10.7551/mitpress/9780262033534.003.0010`.
4. Carlotta Capurro and Gertjan Plets. Europeana, EDM, and the Europeanisation of cultural heritage institutions. *Digital Culture and Society*, 6(2):163–189, 2020. `doi:DOI10.14361/dcs-2020-0209`.

5 Carlotta Capurro, Vera Provatorova, and Evangelos Kanoulas. Experimenting with training a neural network in transkribus to recognise text in a multilingual and multi-authored manuscript collection. *Heritage*, 6(1212):7482–7494, December 2023. `doi:10.3390/heritage6120392`.

6 Sebastian Colutto, Philip Kahle, Hackl Guenter, and Guenter Muehlberger. Transkribus. A Platform for Automated Text Recognition and Searching of Historical Documents. In *2019 15th International Conference on eScience (eScience)*, pages 463–466, San Diego, CA, USA, September 2019. IEEE. `doi:10.1109/eScience.2019.00060`.

7 Read Coop. Training models. data preparation, 2023. URL: `https://help.transkribus.org/data-preparation`.

8 Art DATIS. Art DATIS project. URL: `https://artdatis.nl`.

9 Sven Dupré. Art Educational Reforms and the Studio Glass Movement: Artists' Ambivalent Self-Representation as Actors of Lost Knowledge. In Johan Östling, David Larsson Heidenblad, and Anna Nilsson Hammar, editors, *Knowledge Actors: Revisiting Agency in the History of Knowledge*. Nordic Academic Press, Lund, 2023.

10 Susanne Kaatherine Frantz. *Artists and Glass. A History of International Studio Glass*. PhD thesis, The University of Arizona, 1987.

11 Maya R. Gupta, Nathaniel P. Jacobson, and Eric K. Garcia. Ocr binarization and image pre-processing for searching historical documents. *Pattern Recognit.*, 40:389–397, 2007. URL: `https://api.semanticscholar.org/CorpusID:15012496`, `doi:10.1016/J.PATCOG.2006.04.043`.

12 Stuart Hall. Constituting an archive. *Third Text*, 15(54):89–92, March 2001. `doi:10.1080/09528820108576903`.

13 Karin Hansson, Anna Näslund Dahlgren, and Teresa Cerratto Pargman. Datafication and cultural heritage: Critical perspectives on exhibition and collection practices. *Information I& Culture*, 57(1):1–5, March 2022. `doi:10.7560/IC57101`.

14 Marieke M. A. Hendriksen. Building the ARTECHNE Database: How to Develop a Multi-Purpose Database for an Interdisciplinary Project. *IDEAH*, May 2020. `doi:10.21428/f1f23564.e6779c76`.

15 Anthony Kay. Tesseract: an Open-Source Optical Character Recognition Engine | Linux Journal. *LINUX Journal*, July 2007. URL: `https://www.linuxjournal.com/article/9676`.

16 M.D. Lynn. *American Studio Glass, 1960-1990*. Hudson Hills Press, 2004. URL: `https://books.google.nl/books?id=b4KhNlxnyOEC`.

17 Sue Newell and Marco Marabelli. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification". *Journal of Strategic Information Systems*, 24(2644093), March 2015. `doi:10.2139/ssrn.2644093`.

18 NWO. Smart culture - nwo, February 2022. URL: `https://www.nwo.nl/en/researchprogrammes/creative-industry/smart-culture`.

19 Ada Polak. *Modern Glass*. Faber I& Faber, New York, thomas yoseloff edition, 1962.

20 Vera Provatorova, Samarth Bhargav, Svitlana Vakulenko, and Evangelos Kanoulas. Robustness evaluation of entity disambiguation using prior probes: the case of entity overshadowing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10501–10510, 2021. `doi:10.18653/V1/2021.EMNLP-MAIN.820`.

21 Vera Provatorova, Simone Tedeschi, Svitlana Vakulenko, Roberto Navigli, and Evangelos Kanoulas. Focusing on context is nice: Improving overshadowed entity disambiguation, 2022. `arXiv:2210.06164`, `doi:10.48550/arXiv.2210.06164`.

22 Vera Provatorova, Svitlana Vakulenko, Evangelos Kanoulas, and Johannes M van Hulst. Named entity recognition and linking on historical newspapers: Uva.ilps i& rel at. In *CEUR-WS*, volume 2696, Thessaloniki, 2020. URL: `https://ceur-ws.org/Vol-2696/paper_209.pdf`.

23 RKD. Archief Sybren Valkema. URL: `https://rkd.nl/nl/projecten-en-publicaties/projecten/265-archief-sybren-valkema`.

24 Sybren Valkema and Klaas Laansma. *Sybren Valkema*. De Prom, Baarn, 1994.