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.

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

---

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

---

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.
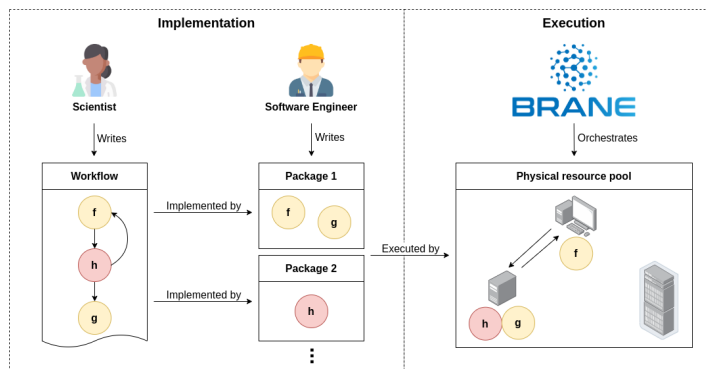

Figure 1 Conceptual role of BRANE. During implementation, a scientist writes a workflow that composes elementary functions (`f`, `g` and `h` in this example). These elementary functions are implemented by packages, which are written beforehand and made publicly available by software engineers. At runtime, 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].
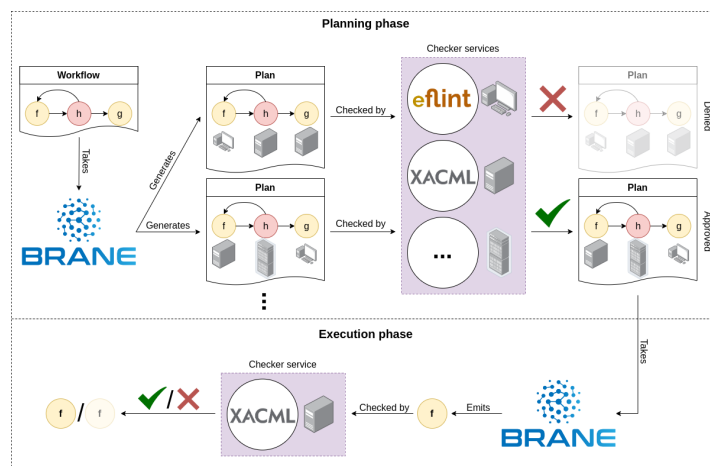
■ **Figure 2 Policy enforcement in BRANE.** There are two phases when executing a workflow: in the first, the *planning phase*, the runtime maps tasks in a workflow to resources that will execute the tasks. Here, checker services (and thus policies) act as filters to separate plans they would allow from 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
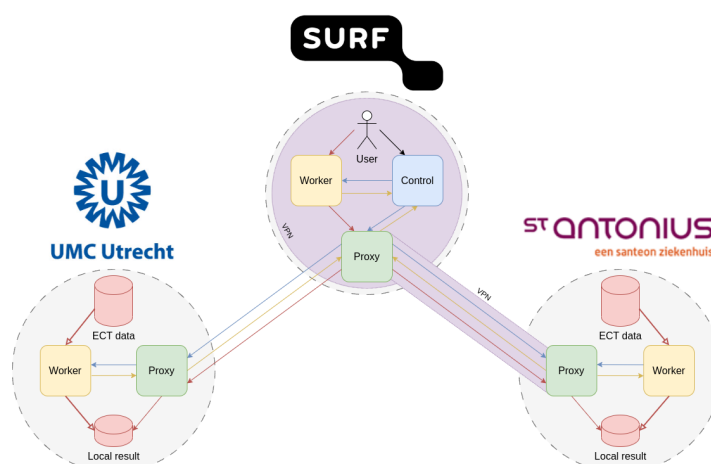
■ **Figure 3 Schematic overview of the PoC setup.** Three domains participate in the proof-of-concept, each of which has their own resources hosting their part of the framework: each hospital (UMC Utrecht and St. Antonius) have local instances of the ECT dataset, and host a worker node to perform local 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 restricted 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`.