

General Learnings From the Horizon 2020 Project BigMedilytics

By Roland Roller, Supriyo Chatterjea, Brian Pickering, Holmer Hemsén, Dimitrios Vogiatzis, Ricard Martínez Martínez, Georg Langs, Simona Rabinovici-Cohen, Wiebke Duettmann, Alex Sangers, Maria-Esther Vidal, Ernestina Menasalvas Ruiz, Marga Martín Sanchez, Josep Redon, Ana Ferrer-Albero, Alexandra Muñoz-Oliver, Gerrit J. Noordergraaf, Igor Paulussen, Per Henrik Vincent, Arne IJpma, José-Ramón Navarro-Cerdán and Santiago Gálvez-Settier

Copyright © 2024 Roland Roller *et al.*

DOI: [10.1561/9781638282372.ch27](https://doi.org/10.1561/9781638282372.ch27)

The work will be available online open access and governed by the Creative Commons “Attribution-Non Commercial” License (CC BY-NC), according to <https://creativecommons.org/licenses/by-nc/4.0/>

Published in *Technology in Healthcare: Introduction, Clinical Impacts, Workflow Improvement, Structuring and Assessment* by Brian Pickering, Roland Roller, Holmer Hemsén, Gerrit J. Noordergraaf, Igor Paulussen and Alyssa Venema (eds.). 2024. ISBN 978-1-63828-236-5. E-ISBN 978-1-63828-237-2.

Suggested citation: Roland Roller, Supriyo Chatterjea, Brian Pickering, Holmer Hemsén, Dimitrios Vogiatzis, Ricard Martínez Martínez, Georg Langs, Simona Rabinovici-Cohen, Wiebke Duettmann, Alex Sangers, Maria-Esther Vidal, Ernestina Menasalvas Ruiz, Marga Martín Sanchez, Josep Redon, Ana Ferrer-Albero, Alexandra Muñoz-Oliver, Gerrit J. Noordergraaf, Igor Paulussen, Per Henrik Vincent, Arne IJpma, José-Ramón Navarro-Cerdán and Santiago Gálvez-Settier. 2024. “General Learnings From the Horizon 2020 Project BigMedilytics” in *Technology in Healthcare: Introduction, Clinical Impacts, Workflow Improvement, Structuring and Assessment*. Edited by Brian Pickering, Roland Roller, Holmer Hemsén, Gerrit J. Noordergraaf, Igor Paulussen and Alyssa Venema. pp. 341–360. Now Publishers. DOI: [10.1561/9781638282372.ch27](https://doi.org/10.1561/9781638282372.ch27).

Big Data, in combination with Artificial Intelligence (AI), has the potential to change and improve processes in medicine. However, these activities/technologies must be developed to promote the trust of all stakeholders: patients, healthcare professionals, private and public providers, and businesses. Providing a trustworthy AI – lawful, ethical, and robust – requires significant efforts. Although technological development is moving quickly, testing, validation, and integration of such innovation may take many years. The reasons that slow down this process are manifold. However, some barriers and pitfalls are foreseeable and, therefore, can be taken into account or avoided. In order to support future development and integration of AI and Big Data technologies, we present technical challenges and lessons learned from our previous project, BigMedilytics, involving clinicians and data scientists. This chapter considers the challenges data scientists providing advanced technology in

the healthcare domain may face, along with some suggestions to address any related issues if applicable.

27.1 Introduction

Sharing experiences can be helpful to make others aware of problems, to learn how to overcome them, and, therefore, to take those difficulties into account. In the context of a technical project, this would enable such projects to plan ahead and save time. This chapter will present possible and common problems and pitfalls while setting up and implementing a Big Data and AI project in the healthcare domain. More specifically, we would like to share the experience of the 12 different BigMedilytics studies. As each study tackled different problems, used different datasets, and dealt with different challenges, the broad variety combined into one project provides much potential to learn from.

Most studies in the project were set up within a hospital to support clinical staff, while only a small number directly targeted patients. The data used in the different studies cover a wide range, including Electronic Medical Records (EMR), clinical text data, images, real-time data from different sources, smartphone data, insurance company claims, biomedical literature, ontologies, and open structured or semi-structured data sources. The studies generally involved a data provider, the custodian of the health-related datasets, and a technical partner processing the data. In most cases, a health register in the community or a hospital represents the data provider, and at the same time, a different external partner carries out the technical implementation. In one instance, data from a hospital were combined using multi-party computation with health insurance claims data, thus, maintaining the security and privacy of the data. Also, in some cases, data providers and technical partners were located in different countries within the EU. Topic-wise, most studies target the prediction of particular outcomes, such as complete pathological response to treatment, risk of cancer recurrence, mortality, risk of hospitalization, infections, exacerbations of Chronic Obstructive Pulmonary Disease (COPD), or heart failure. Others focus on monitoring, for instance, to detect bottlenecks in the usage of particular medical devices, glucose levels, or patient adherence to drug intake. Various studies provide tools to analyse and/or navigate more easily through the given data with the help of AI and Big Data.

Of interest to all stakeholders working on data-driven healthcare propositions, we present the biggest and most crucial technical challenges across the project, along with some lessons learned and solutions. In particular, we discuss the different challenges and consider what would be done differently if we were to do it again. Challenges will be presented, with examples taken from the different studies,

and a possible solution or lessons learned. We present information at different levels, namely: general, data, technical, and validation.

27.2 Prerequisites

Expanding on the general perspectives in response to the internal survey, partners provided specific comments and feedback relating to their studies and their experiences during the project. The main common themes are discussed here.

27.2.1 Interdisciplinary Teams Require Time

Working on Big Data and AI in healthcare involves interdisciplinary work. The stakeholders typically include hospital CEOs, department managers, privacy officers, medical and laboratory staff, system administrators, data scientists, and researchers. This means that people with different educational and professional backgrounds and perspectives need to communicate with each other. It is already difficult to explain the work to a peer when working on complex topics. However, trying to do this with people from totally different backgrounds can lead to miscommunication, frustration, and ultimate failure of the endeavour. Further, bearing in mind that people might use different terminologies for similar things, a language and cultural barrier may exist. Although most people can understand and speak English, they may not appreciate contextual factors or domain-specific jargon. So, it is essential to allocate sufficient time and to have many meetings, particularly at the beginning, to find and then maintain common ground and a common understanding.

27.2.2 Regulatory Protocols that are Incompatible with the Iterative Nature of Scientific Research

In exploratory projects, the clarity on what data are needed to meet a particular objective may evolve over time. There is a fundamental disconnect between regulations and how scientists work. In the hypothetico-deductive tradition, scientists develop a hypothesis, gather the initial data, perform experiments, and derive conclusions that support or question that hypothesis. This might make them realize that they need to collect different data points. In other words, what is needed to address a particular problem may not always be apparent from the start. This is especially true for problems where Big Data is involved.

Different from the financial sector, designing sandboxes in health research is impossible. The sandbox provides a controlled testing environment to enable the

implementation of innovative technology projects. It would mean defining confined environments to conduct data-driven research with the elimination or appropriate mitigation of potential risks. For projects involving partners from several EU Member States or non-EU Members, however, differing legislation must be harmonized and balanced. In addition, effective compliance with GDPR is only possible if it is based on the formal approach. The definition of data protection by design and, by default, legal compliance by design implies a material process appropriate to the conditions of each processing operation and a risk-based approach. The compliance maturity model achieved by GDPR compliance decisions is not theoretical. They involve decision-making by the controller or the processor and generate auditable evidence.

BigMedilytics incorporates valuable lessons learned that inform debate in building the European Health Data Space. At present, legislative asymmetries only allow trans-European research with anonymized data. While many countries exempt consent for retrospective research with data, the requirements for prospective research are diverse. This has the consequence of constraining the design of federated data analytics strategies. In this model, data processing would take place locally, on-premise at a given hospital in a given country, and the results would be shared in the cloud, duly anonymized. In practice, this hampers the possibilities that a European Cloud should provide for data analytics and deploying a common AI strategy for healthcare systems.

27.2.3 Established IT Structures Meet New Requirements

Often the IT infrastructure in hospitals has grown organically over the years and cannot be changed radically due to the need for high availability of services, inherent interdependencies, and external (e.g., government) regulation. While data scientists might be used to powerful computer clusters, Linux machines, and admin rights to quickly install the tools they need, two different worlds collide here. The most significant of these aspects is computational power; the others make the working environment less convenient. However, in cases where the hospital does not provide a powerful enough computer cluster or access is restricted, it may be necessary to buy separate servers, sometimes with GPUs, and integrate them into the existing hospital IT infrastructure. Be aware, this will increase the project (capital) expense, and the integration of new hardware might take time and will doubtless require approval from different departments. Overall, it is essential to be flexible and be able to find quick workarounds to make your system work. In addition, a cloud solution makes the situation easier for the developers. On the other hand, this raises issues related to de-identification, security, and GDPR safeguard. Thus, a hybrid (private) cloud approach might be the optimal solution.

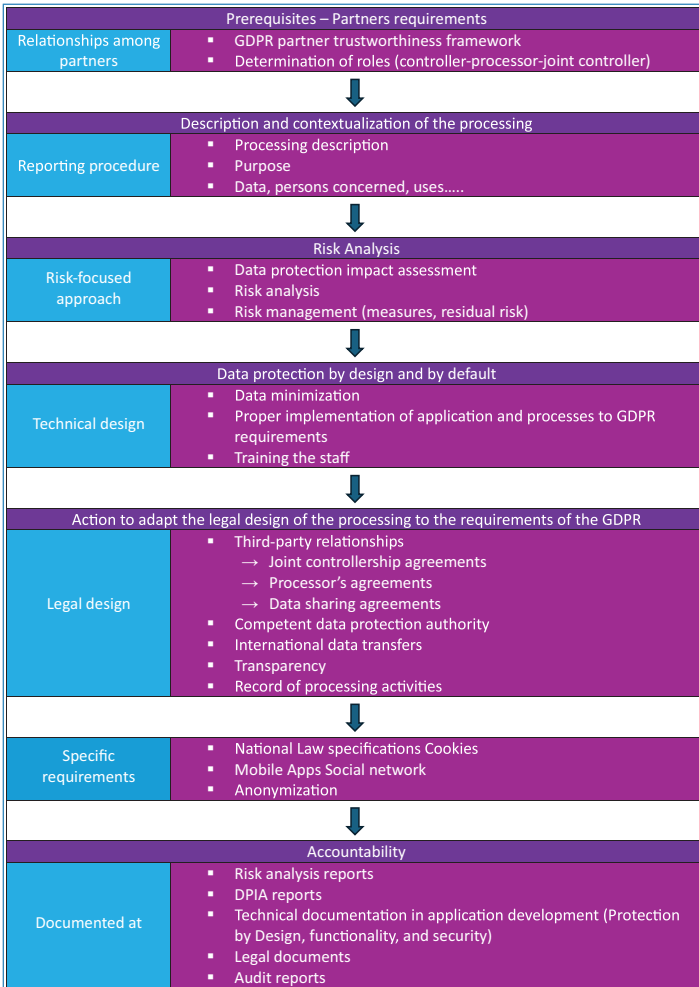


Figure 27.1. Compliance workflow.

27.2.4 New Tools and Clinical Acceptance

Although results might be good within an experimental setup, how can the target group be convinced to use a new model? For instance, certain patient groups may have less experience using apps, such as older or less digitally aware cohorts. Alternatively, clinical staff may be under time pressure and are focused on immediate patient care rather than technological advances. Therefore, they may be reluctant to test or deploy additional tools within their daily routines. At all events, the end user (either patient or clinician) must be convinced of the benefit of a given tool. There are standardized models (e.g., normalization process theory [1]ⁱ),

i. <http://www.normalizationprocess.org/>.

frameworks (e.g., non-adoption, abandonment, scale-up, spread, and sustainability [2]), and programs (e.g., personal and public involvementⁱⁱ) to encourage and facilitate discussion, understanding, and ultimately adoption of new technologies into healthcare contexts. These take time, require significant planning to engage appropriate stakeholders, and may be constrained by existing institution-specific procedures. On the other hand, they depend on good cross-discipline communication and, therefore, can encourage a common language and understanding early on in the project (see above).

From our experience, we achieved good acceptance by preparing introductory material in the form of text or video, including on occasion as part of an app, having direct personal briefings, or including the personnel in the process. Visualization certainly plays an important role. However, it turned out that instead of providing a new tool in addition to all the existing apps and programs, integration into existing working environments, e.g., as an additional feature, might make it easier as medical staff already uses multiple programs daily and may not be open to adding another one so easily unless it is presented as part of existing practices. Finally, trust in the technology might play a crucial role depending on the application. It is important here to see the app as part of a broader sociotechnical context: the technology itself may be robust and completely reliable. However, if the agency promoting its use has lost patient trust, then this will affect take-up negatively.ⁱⁱⁱ Make time to cater to all of these aspects, especially by doing user experiments and engagement. It should be remembered that for healthcare, there are at least two main user groups who need to be collaborated with: first, the clinicians themselves who may have other priorities and may not understand the subtleties of the technologies themselves, and second, the patients who may be suspicious of technologies where they do not see immediate and personal healthcare benefit.

For example, implementing a new study protocol that includes commercial smartwatch technology to track patient activity runs into several security concerns by the data security officer and GDPR compliance officers. Data security officers must contact the commercial provider to validate if proper data security processes are in place. Data of these devices might be stored in the cloud on a different continent, running into GDPR compliance issues. The company selling the commercial devices might be bought by another company during the study, potentially triggering novel GDPR concerns. All these complexities can add significant time delays to a study.

ii. <https://www.health-ni.gov.uk/topics/safety-and-quality-standards/personal-and-public-involvement-ppi>.

iii. This was observed during the COVID-19 pandemic with varying levels of adoption of contact-tracing apps, regardless of each app's reliability, because of suspicion of government or the attitudes of particular groups of users.

27.3 Data

27.3.1 Data Access Across Institutions and/or Countries

Access to data is of fundamental importance to the success of any data-driven initiative. Traditionally, a patient's care has primarily depended solely on the data available to the care provider. However, it is evident that in today's hyperconnected world, the data that could positively impact a patient's health could typically reside across multiple entities and even in multiple countries. Experience gained from the BigMedilytics project has shown that while individual research studies can get access to data after very lengthy procedures, such strategies would not scale in the real world. In fact, even the innovation carried out in research studies would proceed much further if data access mechanisms were more streamlined. For example, in BigMedilytics, there was an instance where a hospital simply could not arrange to have data shared outside its physical boundaries due to privacy/security issues. This prevented it from collaborating with another research institute and resulted in lengthy, drawn-out negotiations. Finally, to resolve this issue, the hospital provided temporary 'visiting researcher' contracts to researchers from the research institute so that they could process the data on the hospital's premises based on its terms and conditions.

Such a construct would be impossible to scale up in the real world and is a clear example of how siloed the world of healthcare is. In fact, the different silos in the healthcare sector are the greatest hurdles that prevent the widescale adoption of Big Data-driven solutions. These silos can exist at different levels: within a hospital, across care providers and other entities (profit/nonprofit organizations), or across countries. Silos within a hospital can be overcome by adopting open platforms that integrate data from different systems. However, for silos beyond the hospital, the technical challenges are less of an issue, and instead, regulations play a more significant role.

In recent years, several techniques for privacy-preserving data analysis (Privacy-Enhancing Technologies or **PETs**), which could circumvent the problems highlighted above, have received much attention. BigMedilytics has focused on one of these techniques, Secure Multi-Party Computation (or **MPC** for short), to demonstrate how sensitive healthcare data can be securely shared and processed across multiple organizations. However, **MPC** (and other privacy-preserving techniques) do not constitute a universal panacea that solves all problems related to data sharing. This is due to several factors, ranging from a technical level – in that designing and implementing an **MPC** solution is far from trivial and often requires more computational power and running time than a conventional solution – to a more legal and societal level, in that jurisprudence on the usage of these techniques is

extremely scarce. Moreover, the exact privacy properties of these techniques often present non-trivial nuances, and ensuring that data owners and controllers properly understand these nuances takes time and effort.

Therefore, there is an urgent need to streamline regulations to improve the competitiveness and innovation potential of the EU at a global level. The following are some points that could help:

- Clearer and updated guidelines (from the European Data Protection Board) on the concept of personal and non-personal data. The EU Member States do not hold a unique and aligned position on the legal concept of personal data (and non-personal data). This limits the capability to re-use health data.
- Clearer and updated guidelines (from the European Data Protection Board) on anonymization techniques. In addition, a code of conduct on anonymization (or anonymization of personal data concerning health) and re-identification risk is also needed.
- Clear guidelines on using privacy-preserving techniques, such as MPC (mentioned above), differential privacy, or federated learning. This point is strongly related to the one above in anonymization, as it often needs to be clarified to what extent these techniques can be seen as forms of anonymization.
- Reduce fragmentation of local conditions on data processing for scientific research purposes, given that member states have leveraged art. 9 (4) GDPR to introduce further limitations to the processing of health data for scientific research purposes, such as the concept of ‘public interest of the research’, the ‘impossibility or disproportionate effort to obtain consent’, or the concept of ‘research institute or body’. This fragmentation limits the capability to process health data in the context of research. In this respect, a code of conduct, followed by harmonizing the local GDPR implementation acts, would be beneficial.
- Reduce fragmentation of local data protection/healthcare rules applicable to health data, particularly in cross-border transfers of health data within the EU.

27.4 Data Access Needs to Comply with Highly Complex Rules and Regulations

As data scientists are not necessarily associated with the data provider, accessing sensitive (special category personal) data in the first place might bring challenges. In our project, some data scientists and data providers were even located in different

countries, which did not make the situation easier. The introduction of the General Data Protection Regulation (GDPR) was intended to harmonize member-state regulation and therefore facilitate well-founded and managed data sharing. However, in practice, it has resulted in many difficulties and further delays.

- Regulation: Periodically, as demonstrated by the pandemic, regulators may announce specific programs to facilitate the sharing of healthcare data.^{iv} It is worth exploring any such opportunities that may apply.
- Governance:
 - Approvals: The sharing of medical data is tightly controlled with oversight, usually from multiple agencies. It is essential, therefore, to begin the ethical approval process as early as possible and especially to be explicit about what data are required and for what purposes.
 - Data curation: Although approval, especially research ethics, will still typically be needed, de-identified or fully anonymous data have reduced risk to the data subject/patient. It is useful, therefore, for members of the team to discuss the appropriateness and impact of fully anonymous data. Further, if data are de-identified or pseudonymized, this should be done by the (clinical) data provider before sharing.
- Technology:
 - Infrastructure: Data are usually protected by special dashboards for computer scientists, which must be programmed, if not already available, or by contracts. Even so, contractual arrangements between medical healthcare providers and guest scientists are difficult to secure and require long processing times. This may also include separate discussions and approvals for any infrastructure to be used to store and process data. Our recommendation would be to engage with a Trusted Research Environment (TRE)^v which conforms with the 5+1 Safes.^{vi}
 - Remote visitation: One approach to this challenge is to use a model-to-data paradigm, where all the data remain within the data provider infrastructure. All computations are applied on a secure server that resides at the data provider premises, and various docker containers and pipelines of analytics models are transferred to the server and executed there. So, if data queries and algorithms are well-defined, the structure of the data they

iv. See, for instance, the COPI Regulations in the UK; <https://www.legislation.gov.uk/ukxi/2002/1438/contents/made>.

v. <https://www.hdruc.ac.uk/wp-content/uploads/2021/04/Goldacre-Review-TREResponse.pdf>.

vi. <https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/fivesafes>.

are to be run against is known. It is worth considering whether the clinical partner (the data provider) can host and run the queries/algorithms the data scientists developed. That way, the raw data are not shared, just controlled access. This requires careful planning and governance but may reduce the administrative burden considerably.^{vii}

- Federation: For a project in which privacy-sensitive data from two or more institutions needs to be combined, privacy-preserving techniques such as MPC (which BigMedilytics used, albeit on synthetic data) offer a potential solution. However, as these are relatively new technologies, conveying the data-security aspect to the respective data security officers is not straightforward. Nevertheless, if multiple datasets are to be used together (collated or cross-correlated), then running a complex query that remotely accesses and temporarily links different data from different sources would again leave the raw data with the data provider and covered by standard operating procedures. Standardized formats and interface 2 are required in this setting as well.

27.4.1 Complexity of Data Rises for Non-experts

Data scientists usually are not medical experts. In addition, real (clinical) data might include many errors (partially due to human input errors, for example, misuse of predetermined fields or use of non-standardized codes) and missing values. Datasets can grow organically over time, and historic design decisions influence the data, but these are not obvious to an ‘outsider’. Thus, in most cases, it is impossible to test methods and directly get excellent and meaningful results. In most cases, there needs to be close interdisciplinary work. Each stakeholder needs a specific understanding of the work of the others in order to achieve satisfying results. For this reason, it is essential to plan frequent technical meetings to share results, foster ongoing and mutual understanding, and ensure that no apparent errors have been made.

Discretizing some variables for anonymization and usability purposes could require different dynamical ranges depending on the considered populations in the study. Moreover, not all clinical measures are obtained at the same time and some of them must be assumed from other heterogeneous past periods related to the dynamics of the corresponding variable. Besides, different and heterogeneous imputation techniques to fill in the empty values for some features in individuals must be applied for a great part of machine learning models to work. The chosen

vii. Care must be taken, of course, that the results of such remote query/execution do not increase the risk of re-identification.

imputation technique depends on some feature properties such as feature nature, probability distribution, and feature relation degree among all related features.

Moreover, the data of the medical domain are of different types and include structured data, text data, genomic data, and imaging of different modalities (e.g., X-ray, MRI, ultrasound, CT, pathology, and so on). Understanding all these modalities and different types of data is complex and requires special expertise. Furthermore, even within the same modality, different medical centres create different data. For example, MRI has no standardized protocol for scan acquisition and high variance of image resolution, voxel size, and image contrast dynamics. This diversity of modalities increases the data complexity and requires special pre-processing and selection of different methods per modality.

27.4.2 Limited Data

While from a medical perspective, a data source may be large, data will likely be too small and with many missing values from a data scientist's perspective. This is because many modern machine learning models are data-hungry. The small data size may introduce biases and not represent the real-world distribution. Also, it significantly decreases the data size if the events you want to detect are seldom. The difficulty of data access for data science in the medical domain is often that the relevant data are distributed across hospitals. This stands in contrast to the majority of data science projects, where the data to analyse is usually either in a single place, can be accessed without restrictions, or is a public source that can be integrated freely. Despite national- or European-level legal regulation for data access, each country and in some countries, even each state and hospital, has its own rules on how data scientists can access and process the data. Moreover, in cases where the goal is to introduce new technology to collect data, e.g., a remote patient monitoring app, work starts from scratch. Beginning to develop methods without data is almost impossible. Where it is necessary to wait until the size of the data increases, rule-based approaches or simple models at the start might help, as well as the generation of some synthetic but representative data. Further, exploiting some additional existing and similar open-access data sources can be beneficial. In such cases, we can either start on that data to develop your first baselines, or blend the data, or pre-train our models.

27.4.3 Data Quality

Data quality in the biomedical domain and clinical care can be critical, as it will inevitably affect patient outcomes, as well as the costs of care. Data quality issues can manifest at multiple steps along a data science pipeline, originating from raw data

but affecting inferred data. To maximize data quality, we recommend supporting standardization at best. For instance, if any clinical staff can enter information freely, the evaluation is complex and requires efforts to standardize afterwards. Furthermore, to standardize diagnoses, we recommend using [SNOMED CT](#); to standardize laboratory values, we recommend using [LOINC](#), and to standardize outcomes, we recommend using [PROMS](#).

Information from biomedical articles or clinical text can support processes and use cases in healthcare. Information about treatments, medications, or adverse drug effects might influence the treatment decision of a caregiver or medical doctor. Thus, methods that extract information should attach a quality or trust score to the extracted information. Regarding the extraction of information from biomedical literature, the publication date of an article, the impact factor of the journal in which it is published, and possibly the authors' affiliations should determine the information's reliability. Investing time in a systematic literature review and a meta-analysis of relevant work may be worth investing time. This should be carried out by experienced personnel.

27.5 Technology

27.5.1 Remote Patient Monitoring

Implementing remote monitoring requires time and patience. We recommend involving all parties (e.g., patients, medical doctors, and nurses, depending on the use case) in the development process (design and such). Standard approaches ([NPT](#), [NASSS](#), [PPIE](#), and so on) have been mentioned above, which would be run in parallel with traditional software engineering processes such as user story analysis. In addition, programming requires time, especially if new features and functionalities need to be implemented. Some extra time should be considered where patients are involved so that the software works well and to agreed standards before the release. This may involve additional testing beyond functional verification. Depending on the use case (e.g., monitoring life-threatening aspects), we do not recommend monitoring patients solely by [AI](#) tools, which certainly would also raise legal concerns. However, for those cases, we suggest putting humans in the loop, e.g., in the form of a telemedicine team.

27.5.2 Image Processing

Analysing medical imaging is generally done via deep neural networks with millions of parameters that need to be learned. Training such a network requires thousands

of image data and some annotations on the images relating to thousands of patients. However, the imaging data available for analytics are scarce and confidential, and access to data is protected and limited. Nevertheless, access to the data for machine learning purposes and permission to display images to radiologists as part of guidelines or as examples can be obtained through approval by an ethics board and suitable anonymization of the images. Moreover, in medical imaging, the annotations require medical expertise and are expensive, time-consuming, and inconsistent. Sometimes multiple modalities are needed as different features are exposed in different modalities. For example, breast density shows up on mammography images but not on ultrasound images, and breast calcifications show up on mammography but typically do not show up via ultrasound and never show up on MRI. Finally, in the medical domain, there is a diversity of populations, genetic variations, and environmental differences that may impact the features exhibited in the imaging. This effect is not quite understood yet. As a result of all these challenges with analysing medical imaging, the creation of robust AI models needs to consider new advanced approaches. Multimodal algorithms that analyse multiple modalities (e.g., CT, MRI, and X-ray), pre-trained models, and transfer learning that reuse models trained on external datasets, as well as federated learning that trains simultaneously on multiple protected datasets, can be beneficial approaches to increase the usable dataset and address the medical imaging AI challenges.

27.5.3 Accessing Information in Text

Much information within Electronic Health Records (EHR) is encoded in semistructured clinical text, such as the well-being of a patient, medication changes, or particular findings. In order to unlock this information, appropriate Natural Language Processing (NLP) tools suited for the clinical domain are required. Overall, we recommend using or building upon existing NLP tools and libraries. Unfortunately, nearly all such tools exist only for English, as are nearly all existing clinical text datasets, which could be used to train a new model. Therefore, working in multilingual Europe on clinical text processing is a major issue and will slow development. While a rule-based approach, such as NegEx for negation detection, can be translated, machine learning-based approaches for more complex problems require labelled training and evaluation data. However, the creation of a new labelled dataset is very time-consuming. Technically, some ways to overcome this challenge are as follows: (1) Research groups working in this field need to publish data or models to contribute to the community. Publishing data, however, is more complicated, as data include sensitive (special category) information, even if de-identified. One solution, for instance, would be merging all de-identified text files

and randomizing the sentences [3]. Publishing models trained on de-identified data might be easier, as models are more abstract. (2) A second possibility lies in modern machine learning techniques such as zero-shot or few-shot learning – training new models on, for instance, similar English data and applying the model to the new target language.

The processing of scientific literature, such as text from PubMed, instead includes, in most cases, English text. However, depending on the use case, the quality of the information provided in publications can be problematic due to outdated articles (information and facts might change over time) or publications from untrustworthy institutes and journals. To this end, when harvesting scientific literature to extract, for instance, related information (e.g., build up knowledge graphs), it is advisable to use specific filters:

- Quality of the journal, check, for instance, h-index or the Scimago Journal Rank (SJR) indicator.
- Filter for particular institutions or authors which are known for their contributions.
- Publication type: Different types of articles are defined according to the different levels of evidence (e.g., scientific review or clinical trial) based on where the represented knowledge is derived.
- Publication year: The recency of a publication allows an expert to decide if the results reported are still relevant.
- The number of citations for a specific publication is a sound measure of its quality and trustworthiness.

27.5.4 Data Quality for Workflow Characterization and Optimization

In order to characterize and improve hospital workflows, hospitals usually only have access to data derived from EMRs. However, while EMRs are excellent for managing patient data, they are not optimally designed to optimize hospital workflows – especially for the ones requiring fine-grained timing information. This is primarily because most data are entered manually in the EMR system. A direct consequence is that data entry is rarely performed exactly at the time a particular action is taken. For example, discharge details of a patient might only be entered into the EMR at the end of a shift. The care provider entering this information thus can only make estimates about the discharge time. Data gathered from BigMedilytics studies have shown that timestamp errors can sometimes be in the order of several hours. To accurately gather timing information, it is essential to understand that many processes within a hospital workflow are closely related to location. For example, in the Emergency Department, the triage, treatment, and discharge processes can be

detected based on the location of a patient. Because of this, data gathered from a Real-Time Locating System (RTLS) can be used to gather accurate timestamps of particular processes automatically. Thus, the RTLS timestamp not only helps to improve the data quality of timestamps but also reduces the burden on staff as the process of entering timestamps can be fully automated.

27.5.5 Strategy to Grow RTLS Infrastructure

Real-time, outdoor location information has radically transformed the way society functions by not just allowing us to locate a position on a map but also by enabling people to perform a wide variety of tasks such as navigating traffic, picking out restaurants, and shopping at a store when it is the least busy. Similarly, real-time indoor location information can transform how healthcare is delivered in hospitals. More specifically, location information from an RTLS infrastructure can significantly improve hospital workflows ranging from asset management to optimizing patient flows. However, an important point to realize is that as most patient and asset trajectories are not limited to a single department but span across multiple departments, any RTLS should ideally be deployed on an enterprise-wide basis.

However, a common misconception is that an enterprise-wide system requires a uniform high-resolution RTLS deployment that can locate any tagged entity down to a room. This is not only expensive but is (in most cases) unnecessary. Instead, a more cost-effective approach is to try and re-use a hospital's existing Wi-Fi infrastructure to act as an RTLS. While a Wi-Fi-based RTLS may only deliver department-level resolution, it does help cover the entire building without investing additional dedicated RTLS infrastructure. Furthermore, once the enterprise-wide Wi-Fi-based RTLS has been rolled out, a hospital can opt to upgrade specific departments or areas that can benefit from more fine-grained location information by using higher-resolution RTLS technologies (such as those based on infrared or Bluetooth). For example, the Emergency Department might be equipped with an infrared-based RTLS to monitor all ED patients or pay special attention to hyperacute (e.g., stroke/sepsis) patients. In addition, Wi-Fi-based RTLS could be used to track ED patients admitted to the hospital and also track mobile assets that move around the hospital. In other words, adopting an open, real-time platform that allows the tagged entities to be seamlessly tracked across multiple high- and low-resolution RTLS technologies is important. This heterogeneous, stepwise approach would allow a hospital to monitor and optimize processes along the entire trajectory of patients while keeping costs in check. Using an open, real-time platform is also a future-proof strategy, as it allows a hospital to build up its capabilities over time and meet its changing needs.

27.5.6 Security

As medical data, by definition, are regarded as a special category of personal data (GDPR, Art 9^{viii}), there are additional requirements on those holding and processing such data to ensure its security. There are standards (e.g., ISO 27001^{ix}), and certification programs (e.g., Cyber Essentials^x; NHS Digital Toolkit^{xi} in the United Kingdom) which provide assurance as to the appropriateness of data storage environments. We recommend that those hosting medical data should explore the options in their context^{xii}: external accreditation of this sort takes some time and may affect budgets. However, they provide an objective indication that the data will be appropriately handled and secure once obtained.

If privacy-preserving solutions such as MPC are used, another challenge rises from the fact that these solutions need to be installed on the IT infrastructure of the data owners, which may pose significant technical and governance challenges, especially given that these solutions often start from a low Technology Readiness Level (TRL) and need to be interfaced with different systems and infrastructures.

27.6 Validation

27.6.1 Comparability

At some point during development, it is important to establish whether the results are sufficient. This may be difficult as development may depend on a single restricted database and even aim to provide insights where no other work has been carried out to date, meaning there are no comparative studies available. Papers on similar tasks, which report results on their data, which can often not be accessed, are only helpful to a small extent, as small differences (task definition, the proportion of positives/negatives, quality/underlying population, and so on) can have a strong influence on the outcomes of your model. A continuing bias exacerbates this situation in the literature to publish only positive findings, not those

viii. Art. 9 GDPR – Processing of special categories of personal data |General Data Protection Regulation (GDPR) (gdpr-info.eu).

ix. <https://www.iso.org/isoiec-27001-information-security.html>.

x. <https://www.gov.uk/government/publications/cyber-essentials-scheme-overview>.

xi. <https://www.dsptoolkit.nhs.uk/>.

xii. Note that, in some cases, such accreditation is essential to process certain datasets. This should be checked as part of planning.

where an approach did not yield useful results. In this regard, we recommend three approaches:

- (1) Try to find a dataset with a similar task and a corresponding benchmark system and test the approach.
- (2) Put sufficient effort into a simple but strong baseline, possibly with the help of domain experts.
- (3) Try to evaluate the system with the end users – although this may be time-consuming. However, the use of systems such as the Observational Medical Outcomes Partnership (OMOP) would solve some of these problems.

27.7 Clinical Validation

In some cases, the AI models can be used in clinical practice only after conducting a clinical trial. AI models that may affect the treatment selection have a direct impact on the patient's health and must be first validated and tested in clinical trials and then approved by regulatory authorities such as the FDA in the United States and the EMA in Europe. This makes clinical validation long and complicated; thus, only a few validation cycles are possible. Additionally, these models need to be interpretable and explainable to increase the acceptance of the AI models. The stakeholders need the ability to interpret the models and understand their reasoning.

27.7.1 Study Design

The study designs should be planned with the help of medical experts and relevant statisticians so that the impact on patients can be evaluated sufficiently and in a way that other medical experts would readily understand and accept the methodology and the findings, leading to their use of technological innovations. It is important to contextualize a given innovation activity within the existing literature and procedures. Clinicians will be used to reading and assessing various trials; they may not so easily follow typical data science publications.

The design of studies involving workflows in hospitals requires some additional considerations. Hospitals are highly dynamic environments. In addition, it may not be possible to control or influence all factors that can impact the outcome of a study. To take these characteristics into account, when executing pre-post studies that focus on evaluating the impact of a particular intervention, it is important not to obtain only two sets of KPI measurements before and after the introduction of the intervention. Instead, tools and procedures should be in place to monitor KPIs continuously at regular periods before and after the introduction of

the intervention. In addition, it is crucial to keep a daily log of all events (e.g., with the help of consultants) that could impact the selected **KPIs**, as the collected information could prove to be critical in retrospectively explaining the characteristics of the **KPIs**. Tools to continuously monitor **KPIs** can also help to check if the introduced intervention is being used correctly or if further training is needed to ensure that the end user (care provider/patient) derives maximum benefit from the solution.

Where multiple partners engage on a project, as with the BigMedilytics trials, ethical approval is likely required from multiple bodies: a relevant health research body and the institute that any academic or data scientist is associated with. Approval should be sought as early as possible and may involve dependencies between different agencies that need to be catered for.^{xiii} Where secondary data are to be used, that is, data collected previously and for another purpose, the data controller or data steward must be consulted to ensure that the data can be used for the proposed trial.

27.7.2 Consent Gathering

The human subjects (e.g., patients) who will participate in a research study need to provide explicit consent before their data can be used for scientific approaches or forwarded to third parties (e.g., data hosts). Thus, the participating medical institutions will need to compose a consent form that requires filling in the name of the patient, the name of the doctor that informs the patient about the study, information about the subject of the study, and the ability to withdraw from the study. The consent form will need to be signed by the human subject. In Germany, for instance, additional consent is required if data are used to establish Big Data and **AI** tools, especially if the data are used by other medical subspecializations (scientists from the radiology field cannot use data from patients with heart diseases). In addition, it is prohibited to use an established prediction model in another context, for instance, in the same patient group but in another hospital. There is the possibility to forward data to third parties (data hosts) if the patient agrees to discard medical privilege in this particular topic (written consent necessary). Better would be to sign contracts with third parties to become a data order processor.

Generally, patients can withdraw their consent at any time without giving a reason. Still, hospitals are advised not to delete data, as they have to provide medical data for at least ten years after production.

^{xiii}. In some countries, for example, the research sponsor will need to approve first. Universities may act as sponsors in this way, but equally, a local health authority may be the sponsor and therefore need to provide approval before the university ethics review board.

Discussing informed consent as a requirement in trials and research studies is common. However, it is important to be clear about what consent is being requested and for what purpose. Briefly, consent may refer to a research participant's agreement to participate in a study, a patient's agreement to undergo treatment, or one legal basis for collecting personal data. By definition, for the consent to be informed, the person giving consent must understand which of these it is. From our experience, we recommend the following:

1. Primary data collection: where you collect data, there are specific requirements:
 - a. A research participant/data subject should be fully informed about the planned purposes, that is, what the data will be used for and who will have access. Make sure, at this stage, that any purposes you are aware of are covered.
 - b. Of course, it may not always be possible to predict how data will be used. It is important, therefore, to let the participant know that future, ethically approved purposes may be found and to give them the option to refuse any such future use, even though they agree to the specific use you have identified.
 - c. Because data are so valuable, it is recommended, wherever possible, to obtain agreement from the research participant for their data to be used, albeit anonymized, in future research.
 - d. Consent should be recorded for audit purposes; research consent does not require a written record.
2. Secondary data use: where you do not collect the data but use data from a different source (an online research database, for instance), then:
 - a. You must check that your intended use of the data complies with the conditions of the data steward.
 - b. You should also check that your intended use of the data is consistent with the original consent provided by the data subject.
 - c. You should make a judgement as to whether the data subject would expect their data to be 'private'. For instance, social media content is not necessarily public domain: there may still be an expectation that content is quasi-private, shared only with trusted others.

Local ethics committees will be able to provide guidance. Most importantly, though, (research) consent should be sought in good time, and any data protection consent could be associated with the potential future assertion of data subject rights (such as withdrawing consent).

27.8 Conclusion

In this document, we presented different challenges along with possible solutions and lessons learned we experienced in a large Big Data and AI project in healthcare. The findings should guide, from a technical perspective, all stakeholders working on data-driven propositions in healthcare.

References

- [1] May CR, Mair F, Finch T, MacFarlane A, Dowrick C, Treweek S, Rapley T, Ballini L, Ong BN, Rogers A, *et al.* Development of a theory of implementation and integration: Normalization process theory. *Implementation Science*. 2009. 4(1), 1–9.
- [2] Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, Hinder S, Fahy N, Procter R, Shaw S, *et al.* Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*. 2017. 19(11), e8775.
- [3] Kittner M, Lamping M, Rieke DT, Götze J, Bajwa B, Jelas I, Rüter G, Hautow H, Sängler M, Habibi M, *et al.* Annotation and initial evaluation of a large annotated german oncological corpus. *JAMIA open*. 2021. 4(2), ooab025.