Lessons Learned from BigMedilytics: Big Data solutions for the Healthcare Sector in Europe

Abstract

Big Data in combination with Artificial Intelligence (AI) has the potential to change and improve processes in medicine. These activities/technologies must be developed in a way that promotes the trust of all stakeholders: patients, healthcare professionals, health care private and public providers and business. Providing a Trustworthy AI, -lawful, ethical and robust-, requires significant efforts for all. Although technological development is moving quickly, test, validation and integration of such innovations may still take many years. Reasons which 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 learnt from BigMedilytics, a large lighthouse project involving both clinicians and data scientists. This document aims at sharing the main findings of data scientists of the project to contribute to the community.

1. Introduction

Sharing experience can be helpful to make others aware of problems, to learn how to overcome them and therefore to take those difficulties into account and plan ahead. In the context of a technical project, this can mean to proceed quicker and therefore to save time. In this document we would like to 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 a large innovation action project, BigMedilytics (Big Data for Medical Analytics), which was a lighthouse project funded by the European Commission from 2018-2021. It consists of 12 different pilots that cover three themes: Population Health and Chronic Disease Management, Oncology and the Industrialization of Healthcare Services. As each pilot tackled different problems, used different datasets and dealt with different challenges, the large number of varieties combined into one project bears much potential to learn from.

Most pilots of the project are set up within a hospital to support clinical staff (#11), while only a small number (also) target patients directly (2). Most frequently used data sources are electronic medical records (EMR) (7), clinical text data (5), images (3), real time data from different sources (4), smartphone data (5), insurance company claims (1), biomedical literature (1), ontologies (1), and open structured or semi-structured data sources (1). In most cases a health register in the community or a hospital represents the data provider (X), and at the same time, a different, external partner (X) carries out the technical implementation. In one instance data from a hospital was combined using Multi Party Computation with Health insurance claims data, thus maintaining security and privacy of the data. Also in various cases, data providers and technical partners were located in different countries (X). Topic-wise most pilots target the prediction of particular outcomes (X), such as pathological complete response to treatment, risk of cancer recurrence, mortality, risk of hospitalization, infections, exacerbations of COPD or heart failure. Others focus on the aspect of monitoring, for instance, to detect bottlenecks in the usage of particular medical devices, glucose levels, or the adherence of drug intake of patients. A variety of other pilots provide tools to analyse and/or to navigate more easily through the given data with the help of AI and Big Data.

Addressed to all stakeholders working on data driven propositions in healthcare, we present in the following the biggest and crucial technical challenges across the project, along with some lessons learned and solutions. In particular, we discuss the different challenges and take into consideration, what would be done differently, if we do it again. Challenges will be presented, together with examples taken from the different pilots, and a possible solution or lesson learned. We present information at different levels, namely: general, data, technical and validation.

2. Challenges and Lessons Learned

In collaboration with WP5, we developed and ran a small survey to identify challenges and issues that partners thought important as we came towards the end of the project. The survey was generated using assertions agreed by experts in a previous Delphi study[1] and with reference to some work we had done on informed consent in relation to advanced technologies[2]. The assertions were grouped into four categories as follows:

Table 1: Main categories covered by individual assertions

Category	**Main Assertion and Number of Individual Assertions**	
Requirements	What are stakeholders' expectations from advanced technologies?	8
Design & Responsibility	How should advanced technologies be designed?	7
Ethics & Governance	How should advanced technologies be managed?	8
Transparency	How should advanced technologies operate?	7
	TOTAL	30

Participants were asked to record their agreement with each of the 30 assertions across the four categories on a four-point Likert scale (Strongly agree, Agree, Disagree, Strongly disagree). The most significant results are summarised below.

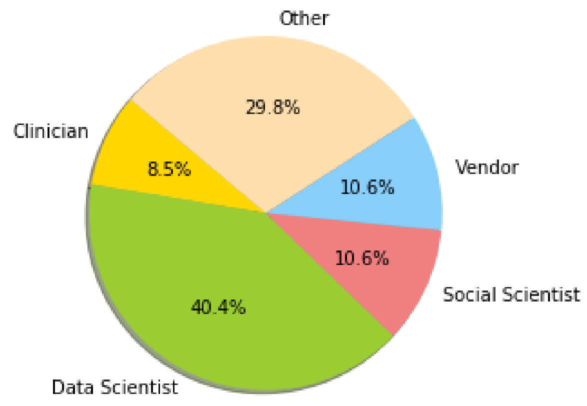
There were 47 participants in total, providing 46 consistent responses; it took on average 14m11s to respond. Job roles covered were weighted towards non-clinicians (see Figure 1); but domains were reasonably well covered[3] (see Figure 2).

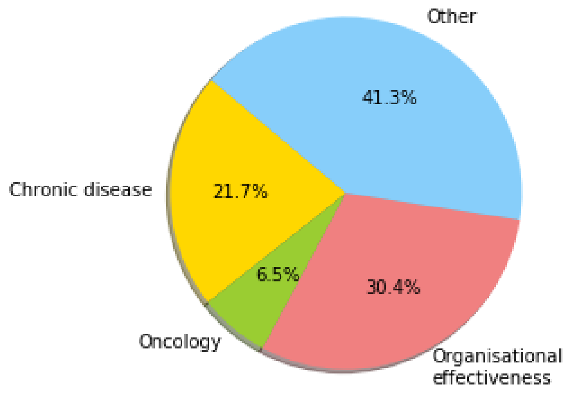
Figure 1: Job Roles of Participants	Figure 2: Domain of Participants

[1] https://doi.org/10.5281/zenodo.1303252

[2] https://doi.org/10.3390/fi13050132

[3] Oncology *per se* seems to be under-represented, though different pilots may involve cancer diagnosis and cancer treatment.





In general, participants agreed with the assertions from the experts in the Taylor et al. (2018) Delphi study about Responsible AI. However, using:

(Strongly agree + Agree) < (Disagree + Strongly disagree)

as an approximate measure of disagreement, the following concerns were raised (see Table 2).

Table 2: Domain of Participants

Assertion	**Disagreement**	**Conclusion**
I depend on technology to do my job, but I'm not responsible for the results that come from the technology	34/47	Relevant actors recognise that everyone has some level of responsibility for the technology they use.
RECOMMENDATION: where advanced technologies are to be deployed, all main actors (those directly involved with the technology) and all other stakeholders (those affected by the technology) should be consulted.		
So long as the technology works, we don't need to worry about ethics	44/47	Actors recognise that there are ethical standards to be met.
RECOMMENDATION: advanced technology testing should include an ethics audit along with standard testing		
A government seal of approval would be enough for people to trust advanced technologies	41/47	Actors do not necessarily trust official accreditation schemes.
RECOMMENDATION: all actors (and stakeholders) need some visibility and oversight of advanced technology deployment; it's not enough to have a separate certification authority		

If appropriate, participants were encouraged to leave free-form comments. There were five comments in total, two identified discomfort about trying to decide whether they really did agree or not (i.e., procedural issues common in such surveys). The other three were:

1. *“Technicians and 'other people' need to find or develop a common language to be able to discuss the pro's and con's of AI.”* This highlights the need for different disciplines to collaborate on the basis of a shared understanding (“common language”)

2. *“Advancing technology may become a new field in which multi-disciplines work together. (the Big Data project is potentially an example of techno-clinical collaboration...).”* Taking the perspective of collaboration from the previous comment forward, this recognises the importance of projects like BigMedilytics to encourage cross-disciplinary work, and, of course, to share experience.

3. *“We need to start viewing the world as a socio-technical system where humans and technologies are networked together and inseparable [from] each other.”* This comment highlights the complexity of the ecosystem around and dependent on advanced technologies. It is essential (as highlighted in the survey responses themselves) to rethink how all actors and stakeholders need to be and can be involved, or at least be appropriately represented.

2.1. General

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

Interdisciplinary teams require time Working on Big Data and AI in healthcare should include interdisciplinary work. This was highlighted by our internal survey (see above). 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 different perspectives need to communicate with each other. Working on complex topics, it might already be difficult to explain work to a peer. But trying to do this with people from a totally different background can lead to miscommunication, frustration and ultimate failure of the endeavour. Further, bearing in mind that people might use a different terminology for similar things anyway, a language and a cultural barrier may exist. Although most people are able to 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 in order to find and then maintain a common ground.

Regulatory protocols that are incompatible with iterative nature of scientific research. In exploratory projects, the clarity on what data is needed to meet a particular objective could evolve over time. There is a disconnect between regulations and how scientists fundamentally work. Scientists develop a hypothesis, gather the initial data, perform experiments and derive conclusions that 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 right from the start. This is especially true for problems where Big Data is involved.

Unlike in the financial sector, it is not possible to design sandboxes in health research. The sandbox provides a controlled testing environment to enable the implementation of innovative technology projects. It would mean defining confined environments in which to conduct data-driven research with the elimination or appropriate mitigation of potential risks. For projects involving partners from several EU Member States, however, differing legislation has to be harmonised and balanced. In addition, effective compliance with GDPR is simply impossible if it is based on the mere formal approach. The definition of data protection by design and by default, in fact the legal compliance by design, implies a material approach to the conditions of each processing operation and a risk-based approach. In the compliance maturity model achieved by GDPR compliance decisions are not theoretical, they involve decision making by the Controller or the Processor and generate auditable evidence.

Table 3. Compliance workflow.

Prerequisites – Partners requirements	
Relationships among partners	▪ GDPR partner trustworthiness framework ▪ Determination of roles (controller-processor-joint controller)
Description and contextualization of the processing	
Reporting procedure	▪ Processing description ▪ Purpose ▪ Data, persons concerned, uses.....
Risk Analysis	
Risk-focused approach	▪ Data protection impact assessment ▪ Risk analysis ▪ Risk management (measures, residual risk)
Data protection by design and by default	
Technical design	▪ Data minimization ▪ Proper implementation of applications and processes to GDPR requirements ▪ Training the staff
Actions to adapt the legal design of the processing to the requirements of the GDPR	
Legal design	▪ Third-party relationships. → Joint controllership agreements → Processor's agreements → Data sharing agreements ▪ Competent data protection authority ▪ International data transfers ▪ Transparency ▪ Record of processing activities
Specific requirements	▪ National Law specifications.Cookies. ▪ Mobile Apps Social networks ▪ Anonymization
Accountability	
Documented at	▪ Risk analysis reports ▪ DPIA reports. ▪ Technical documentation in application development (PbD, functionality, security)

	▪ Legal documents. ▪ Audit reports

BigMedilytics incorporates valuable lessons learned that might inspire debate in building the European Health Data Space. At present, legislative asymmetries only allow trans-European research with anonymised data. While many countries exempt consent for retrospective research with data, the requirements for prospective research are very diverse. This has the consequence of forcing the design of federated data analytics strategies. In this type of model, data processing would take place locally, in the computer premises of a given hospital in a given country, and the results would be shared in the cloud, duly anonymised. In practice, this hampers the possibilities that a European Cloud should provide for data analytics and the deployment of a common Artificial Intelligence strategy for Healthcare Systems.

Established IT-structures meet new requirements Often the IT-infrastructure in hospitals has grown organically over 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 tools they need, two different worlds collide here. The most significant of these aspects is probably computational power; the others simply make the working environment less convenient. However, in cases where the hospital does not provide a powerful enough computer cluster or access is restricted in any way, it may be necessary to buy separate servers, sometimes with GPUs, and integrate it into the existing IT-infrastructure. Be aware, this will increase the project (capital) expense, the integration of new hardware might take time, and will doubtless require approval from different departments. Overall, it is important to be flexible and be able to find quick workarounds to make your system work.

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 with using apps in general, such as older or less digitally aware cohorts. Alternatively, clinical staff may be under time pressures and are focused on immediate patient care rather than technological advances, and may therefore be reluctant to test or deploy additional tools within their daily routines. At all events, the end user (either patient or clinician) needs to be convinced of the benefit of a given tool. There are standardised models (e.g., Normalisation Process Theory[4]), frameworks (e.g., Non-adoption, Abandonment, Scale-up, Spread, Sustainability[5]) and programs (e.g., Personal and Public Involvement[6]) 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 well be constrained by existing institution-specific procedures.

[4] NPT: http://www.normalizationprocess.org/; https://doi.org/10.1186/1748-5908-4-29

[5] NASSS: https://doi.org/10.2196/jmir.8775

[6] PPI: https://www.health-ni.gov.uk/topics/safety-and-quality-standards/personal-and-public-involvement-ppi

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 into 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 add another one so easily unless it is presented as part of existing practices. Finally, depending on your application, trust in the technology might play a crucial role. It is important here to see the app as part of a broader socio-technical context: the technology itself may be robust and completely reliable. However, if the agency promoting its use have lost patient trust, then this will affect take-up negatively[7]. Make time to cater for all of these aspects, especially by doing some user experiments and engagement. It should be remembered that for healthcare there are 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 secondly, patients who may be suspicious of technologies where they don't see immediate and personal healthcare benefit.

For instance, implementing a new study protocol that includes the use of commercial smart watch technology to track patient activity runs into several levels of security concerns by the data security officer and GDPR compliance officers. Data security officers will have to 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 regulations. 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 delay to a study.

2.2. Data

Data access across institutes and/or countries. Access to data is of fundamental importance to the success of any data-driven initiative. Traditionally, the care of a patient has primarily been dependent solely on the data available at the care provider. However, it is evident that in today's hyper-connected world, the data that could positively impact a patients' health could typically reside across multiple entities and even in multiple countries. Experience gained from the BigMedilytics project has shown that while individual research pilots may be able to get access to data after very lengthy procedures, in the real-world, such strategies would not scale. In fact, even the innovation carried out in research pilots would proceed much further if the 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 long drawn 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 own terms and conditions.

[7] This was seen 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.

Such a construct would obviously 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 which prevent the wide-scale adoption of Big Data driven solutions. These silos can exist at different levels: within a hospital, across care providers and other entities (profit/non-profit organizations) or across countries. Silos within a hospital can be overcome through the adoption of open platforms that allow data from different systems to be integrated. However, for silos beyond the hospital, the technical challenges are less of an issue and instead regulations play a greater role.

In recent years, several techniques for privacy-preserving data analysis (Privacy-enhancing technologies or PETs), which could in principle circumvent the problems highlighted above, have gathered a lot of attention. BigMedilytics has focused on one of these techniques, known as 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 for being trivial and often require 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 data controllers properly understand these nuances is a time-consuming process.

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 data 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) is also needed.

- Clear guidelines on the usage of 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 is often unclear 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 a harmonisation of the local GDPR implementation acts would be very helpful.

- Reduce fragmentation of local data protection/healthcare rules applicable to health data, in particular in the field of cross-border transfers of health data within the EU.

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 well 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 General Data Protection Regulations (GDPR) was intended to harmonise member-state regulation and therefore facilitate well-founded and managed data sharing. In practice, though, it 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[8]. It is worth exploring any such opportunities which 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 is required and for what purposes.
    - **Data curation:** although approval 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 pseudonymised, this should be done by the (clinical) data provider before sharing.
- **Technology:**
    - **Infrastructure:** Data is 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 not easy to secure and require long processing times. This may also include separate discussion 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)[9] which conforms with the 5+1 Safes[10].
    - **Remote visitation:** One approach to this challenge is to use a model-to-data paradigm where all the data remains at 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 and the structure of the data they are to be run against is

[8] See, for instance, the COPI Regulations in the UK; https://www.legislation.gov.uk/uksi/2002/1438/contents/made

[9] https://www.hdruk.ac.uk/wp-content/uploads/2021/04/Goldacre-Review-TRE-Response.pdf

[10] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/access-control/five-safes

known, then it is worth considering whether the clinical partner (the data provider) can host and run the queries / algorithms which the data scientists developed. That way the raw data is not shared, just controlled access to it. This requires careful planning and governance but may reduce the administrative burden considerably[11].

- **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, since 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 which 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 interfaces[2] are required in this setting as well.

Complexity of data rises for non-experts Data scientists are normally 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 not possible to just test your methods and directly get good and meaningful results. In most cases there needs to be close interdisciplinary work. Each stakeholder needs a certain 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 obvious errors have been made.

Moreover, the data of the medical domain is of different types and includes structured data, text data, genomic data, imaging of different modalities (Xray, MRI, Ultrasound, CT, pathology, and more). Understanding all these modalities and different types of data is complex and requires special expertise. Even within the same modality, different medical centers 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 selecting different methods per modality.

Limited data While from a medical perspective, a data source might be large, data will be most likely too small and with many missing values from a data scientist perspective. This is due to the fact that 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 size of your data, if the events you might want to detect are seldom. The difficulties of data access for data science in the medical domain is often that the relevant data is distributed across hospitals. This stands in contrast to the majority of data science projects, where the data to analyse is usually either at a single place, can be accessed

[11] Care must be taken, of course, that the results of such remote query / execution does not itself increase the risk of re-identification.

without restrictions or is a public source which can be integrated freely. Despite national or European-level legal regulation for data access, each country and in some countries even each state as well as hospital has its own rules on how data scientists can get access and process the data. Moreover, in cases where the goal is to introduce a new technology to collect data, e.g., a remote patient monitoring app, work actually 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, you can either start on that data to develop your first baselines, or blend the data, or pre-train your models.

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 maximise data quality, we recommend supporting standardization at best. For instance, if any clinical staff can enter information freely, the evaluation is difficult and requires efforts to standardize afterwards. Furthermore, in order to standardize diagnoses, we recommend the use of SNOMED CT; to standardize laboratory values, we recommend the use of LOINC; and to standardize outcomes, we recommend the use of 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 extracts information should attach a quality or trust score on the extracted information. Regarding the extraction of information from biomedical literature, also publication date of an article, the impact factor of the journal in which it is published, and possibly the affiliations of the authors should determine the reliability of the information. It may be worth investing time for a systematic literature review and a meta-analysis of relevant work. This should be carried out by experienced personnel.

2.3. Technology

Remote Patient Monitoring To implement 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, PPI etc.) have been mentioned above which would be run in parallel with traditional software engineering processes such as user story analysis, and so forth. 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 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.

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 generally requires thousands of image data and some annotations on the images relating to thousands of patients. However, the imaging data available for analytics is scarce, confidential and access to it is protected and limited. Nevertheless, access to the data for machine learning purposes as well as permission to display images to radiologists as part of guidelines or as examples can be obtained through approval by an ethics board as well as suitable anonymization of the images. Moreover, in medical imaging, the annotations require medical expertise, 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, breast calcifications show up on mammography but typically don’t 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 have an impact on the features exhibited in the imaging, and this effect is not quite understood yet. As a result of all these challenges with analysing medical imaging, creation of robust AI models needs to consider new advanced approaches. Multimodal algorithms that analyse multiple modalities (e.g. CT, MRI, XRay), pre-trained models and transfer learning that reuse models trained on external datasets, and 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.

Accessing information in text Much information within Electronic Health Records (EHR) is encoded in semi-structured clinical text, such as well-being of patient, medication changes or particular findings. In order to unlock this information, appropriate NLP (natural language processing) tools, suited for the clinical domain, are required. However, nearly all such tools exist only for English, as are nearly all existing clinical text datasets, which could be possibly used to train a new model. Therefore, working in multilingual Europe on clinical text processing is a major issue and will certainly slow down the development. While a rule-based approach, such as NegEx for negation detection, can possibly be simply translated, machine learning based approaches for more complex problems require labelled training, as well as evaluation data. Particularly the creation of a new labelled data set is very time consuming. Technically, there are some ways to overcome this challenge: 1) Research groups working in this field need to publish data or models to contribute to the community. Publishing data, however, is more difficult, as data includes sensitive (special category) information, even if de-identified. One solution, for instance, would be merging all de-identified text files and randomising the sentences[12]. Publishing models trained on de-identified data might be an easier solution as models are more abstract. 2) A second possibility lies in modern machine learning techniques such as zero-shot or few-hot learning. Training new models on for instance similar English data, and applying the model to the new target language.

Data quality for workflow characterisation and optimization In order to characterise 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 for optimizing hospital workflows – especially for the ones where fine grained timing information is required. This is primarily because most of the data is entered manually in the EMR system. A direct consequence is that data entry is rarely performed exactly at the

[12] https://academic.oup.com/jamiaopen/article/4/2/ooab025/6236337?login=true

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 pilots has shown that errors in timestamps can be in the order of several hours in certain cases. To accurately gather timing information it is important 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 clearly detected based on the location of a patient. In view of this, data gathered from a Real-Time Locating System (RTLS) can be used to automatically gather accurate timestamps of particular processes. Thus, the RTLS time stamp not only helps to improve data quality of timestamps but also reduces burden on staff as the process of entering timestamps can be fully automated.

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 navigate traffic, pick out restaurants and shop at a store when it is the least busy. Similarly, real-time indoor location information has the capability to transform the way healthcare is delivered in hospitals. More specifically, location information from an RTLS infrastructure can play a significant role in improving various hospital workflows ranging from asset management to optimizing patient flows. 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 WiFi infrastructure to act as an RTLS. While a WiFi-based RTLS may only deliver department-level resolution, it does help cover the entire building without having to invest additional dedicated RTLS infrastructure. Once the enterprise wide WiFi-based RTLS has been rolled out, a hospital can opt to upgrade certain 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, a WiFi-based RTLS could be used to track ED patients who are admitted to the hospital and also track mobile assets which move around the hospital. In other words, it is important to adopt an open, real-time platform that allows the tagged entities to be seamlessly tracked across multiple high and low-resolution RTLS technologies. This heterogeneous, stepwise approach would allow a hospital to monitor and optimize processes along the entire trajectory of patients while keeping costs in check. The use of 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.

Security Clearly, since medical data by definition is regarded as special category personal data (GDPR, Art 9[13]), there are additional requirements on those holding and processing such data to ensure its security. There are standards (e.g., ISO 27001[14]), and certification programs (e.g., Cyber Essentials[15]; NHS Digital Toolkit[16], in the UK) 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[17]: external accreditation of this sort takes some time and may affect budgets. However, once obtained, they provide an objective indication that the data will be appropriately handled and secure.

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 significative 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.

2.4. Validation

Comparability At some point during development, it's important to be able to establish whether the results obtained are sufficient. This may be difficult since 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 cannot often not be accessed, are only helpful to a small extent, as small differences (task definition; proportion of positives/negatives; quality/underlying population etc.) can have a strong influence on the outcomes of your model. This situation is exacerbated by a continuing bias in the literature to publish only positive findings and not those where an approach did not return apparently useful results. In this regard, we recommend three different approaches: 1) Try to find a dataset with a similar task and a corresponding benchmark system and test your approach in that way. 2) Put sufficient effort into a simple, but strong baseline, possibly with the help of domain experts. 3) Try to evaluate your system with the end users - although this may turn out to be too time-consuming. However, the use of systems such the Observational Medical Outcomes Partnership (OMOP) would solve some of these problems.

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 direct impact on the patient's health, and must be first validated and tested in clinical trials, and then approved by the regulatory authorities such as the FDA in the US and the EMA in Europe. This makes the clinical validation long and difficult, and thus only few validation cycles are possible. Additionally, to increase the acceptance of the AI models, these models

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

[14] https://www.iso.org/isoiec-27001-information-security.html

[15] https://www.gov.uk/government/publications/cyber-essentials-scheme-overview

[16] https://www.dsptoolkit.nhs.uk/

[17] Note that in some cases such accreditation is essential to be able to process certain datasets. This should be checked as part of planning.

need to be interpretable and explainable. The stakeholders need the ability to interpret the models and understand their reasoning.

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 contextualise a given innovation activity within the existing literature and procedures. Clinicians will be used to reading and assessing various types of trials; they may not so easily follow typical data science publications.

The design of studies involving workflows in hospitals require 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 which focus on evaluating the impact of a particular intervention, it is important to not obtain only two sets of KPI measurements before and after the introduction of the intervention. Instead, tools and procedures should be in place to continuously monitor KPIs at regular time periods both before and after the introduction of the intervention. In addition, it is important 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 properly 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 is the case with the BigMedilytics trials, ethical approval is likely to be 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 which need to be catered for[18]. Where secondary data is to be used, that is data collected previously and for another purpose, then the data controller or data steward must be consulted to ensure that the data can be used for the proposed trial.

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 institute will need to compose a consent form that requires to fill-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 as well the ability to withdraw from the study. The consent form will need to be signed by the human subject. In Germany for instance, an additional consent is required if data is used to establish Big Data and AI tools. Especially, if the data is used by other medical sub-specializations (data from patients with heart diseases cannot be used by scientists from the radiology field). In addition, it is prohibited to use an established prediction

[18] In some countries, for example, the research sponsor will need to give approval 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.

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

It is common to talk about informed consent as a requirement in trials and research studies. However, it is important to be clear what consent is being requested and for what purpose[19]. Briefly, consent may refer to a research participant's agreement to take part 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, then 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 is so valuable, it is recommended wherever possible to obtain agreement from the research participant for their data to be used, albeit anonymised, 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

[19] *See Pickering, 2021, in footnote 2, developed in collaboration with a Turing-funded project and which discusses different types of consent.*

associated with the potential future assertion of data subject rights (such as withdrawing consent).

Conclusion

In this document, we presented different challenges along with possible solutions and lessons learnt we experienced in a large Big Data and AI project in healthcare. The findings should provide some useful guidance from a technical perspective for all stakeholder working on data driven propositions in healthcare.