Biomedical Alliance in Europe

Case Studies from medical societies on the European Health Data Space

The Biomedical Alliance in Europe (BioMed Alliance), a unique initiative of 36 medical and research societies, has followed the legislative discussions on the European Health Data Space (EHDS) proposal with great interest. We very much welcome the initiative to facilitate health data sharing for healthcare and research, as we believe both will generate significant added value for patients.

Healthcare professionals and researchers represented within BioMed Alliance believe that we should reduce barriers to health data sharing, while putting in place the appropriate safeguards and maintaining a harmonised approach across the EU. Members of our alliance have collected a set of examples of the secondary use of data in health research and highlight the challenges they face.

Experiences show that the European Health Data Space legislation needs to take into account the experiences of researchers and facilitate health data sharing. While BioMed Alliance believes that the original European Commission Proposal on EHDS included sufficient safeguards to protect patient privacy, we recognise that an opt-out approach could be an appropriate compromise. An opt-out mechanism needs to apply to certain conditions not to hinder life saving health research, and needs to be implemented in a way that prevents data on certain patient groups from being excluded, as this could mean that important insights on their health will not be generated (read our statements here and here).

Medically assisted reproduction[1]

In the field of medically assisted reproduction, patients and professionals need estimations of the chances of a treatment resulting in a live birth to inform treatment decisions. These estimations require data that allow calculating cumulative outcomes across treatment cycles, which is only possible if the data are at least pseudonymized, so that cycles from the same patient can be linked.

Prostate cancer screening[2]

As the European Recommendations on Cancer Screening currently recommend pilots and further research on risk-based Prostate Cancer screening, it will be important for the EHDS to facilitate comparison and research across datasets and countries to allow researchers to understand more about which groups of men need which follow up. In order to develop risk-based screening guidelines, public health and urological researchers need to access pseudonymised data to be able to understand more about which men have the highest risk of developing Prostate Cancer that can have serious impact on their lives. This cannot be accomplished with anonymized data – researchers need to know age, background, ethnicity, etc, and also need to link screening data with patient registries to monitor outcomes, which all help to guide better decision making. For Prostate Cancer

[1] Case study provided by ESHRE
[2] Case study provided by EAU

in particular we know men with family history or the BRCA2 gene mutations are more at risk, as are Black men (who have twice the risk of developing Prostate Cancer). Any option which could introduce a bias to population data means that clinical guidelines do not accurately reflect the needs of people who are not captured in the data, thus screening and care for these groups is sub-optimal.

Cardiology[3]

Opt – out has been found to be a reasonable compromise between respecting the patients' preferences and accuracy of the gathered data, whereas patients with neurologic communication deficits and impaired capacity to consent were proven to be severely underrepresented within opt-in regimens in a 2022 study.[4]

In addition, there is a risk that certain patient groups are excluded and there is risk of incomplete datasets and a lack of information on certain population groups. An algorithm used to estimate healthcare needs in the US severely underestimated the medical needs of black patients due to a bias in the selection of the data: in order to predict the healthcare need the data used health care costs as a proxy, but given that the access to care is uneven to black people, their needs were severely underestimated. What followed was a reduction of black people receiving medical care estimated between 18 % to 47 %[5].

Black and Hispanic women in the United States tend to experience heart-related health issues earlier than White women. Nevertheless, the overall average data for the onset of these diseases is considerably higher, and this means that the medical needs of Black and Hispanic are severely underestimated and not properly addressed. For example, a 2022 study revealed that while a condition called ISH typically starts around age 60 for women in the US, its average onset was 49.3 years and even earlier for Black women, between 41.9 to 46.4 years, and the cause for the discrepancy and hence the disparity in treatment and care is to be found in the selection bias within the dataset used.[6]

Commented [MM1]: I think maybe we can delete this paragraph

A 2016 study showed that opt-out regimens substantially increased HIV testing, while opt-in regimens reduced testing. The study provided evidence that small changes in wording can significantly affect patients' behaviour and thus the understanding of their preferences. Specifically, modifying HIV testing defaults led to clinically and statistically significant differences in test

[3] Case study provided by ESC

[4] Napierkowski, I., Lorenz-Meyer, I., Hille, A., Ebinger, M., Freitag, E., Harmel, P., Endres, M., Hagemann, G., Koennecke, H., Mackert, B., Siegerink, B., & Audebert, H. J. (**2022**). **Follow-up of patients with stroke based on opt-out choice**. *Neurology*, 99(13), e1335–e1344. https://doi.org/10.1212/wnl.0000000000200916 https://pubmed.ncbi.nlm.nih.gov/35918161/

[5] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (**2019**). **Dissecting racial bias in an algorithm used to manage the health of populations**. *Science*, 366(6464), 447–453. https://doi.org/10.1126/science.aax2342 https://www.science.org/doi/10.1126/science.aax2342

[6] Reeves, A. N., Elliott, M. R., Lewis, T. T., Karvonen-Gutierrez, C., Herman, W. H., & Harlow, S. D. (**2022**). **Study selection bias and racial or ethnic disparities in estimated age at onset of cardiometabolic disease among midlife women in the US**. *JAMA Network Open*, 5(11), e2240665. https://doi.org/10.1001/jamanetworkopen.2022.40665

acceptance percentages, whereas opt-out test offers were accepted 28 percentage points more often than opt-in offers[7].

Commented [MM2]: Not sure this example is clear enough

Radiology[8]

Commented [MM3]: Is very long but I quite like it, maybe see if we can cut

Most (especially three-dimensional) radiological imaging is virtually impossible to anonymize. The level of depicted anatomical details in modern imaging (e.g., size and location of organs and bones, geometrical details of vessels, location of cysts in the kidneys or the pancreas, details of brain sulci etc.) will in many cases allow for reidentification if the same person was scanned more than once[9].

As for many research purposes the original patient data itself does not necessarily need transfer, privacy could in many cases be guaranteed by design (e.g., using federated learning approaches for development of AI / independent local validation of models on-site with only aggregated results being shared).

The facilitation of secondary use of healthcare data for research will be essential for the future evolution of science and healthcare in general towards more data-based approaches. Even with the use-cases below much of what would be possible if pseudonymized data was accessible at a large scale (which seems only possible with opt-out approaches) is yet unknown. Especially with the goal of evolving to more predictive health care the need for inclusive data and access to this data is essential to develop health policy at the European level. Not enabling the most comprehensive datasets possible, while preserving patient privacy and security, would be inconsistent with the goals of the EHDS in our opinion.

Need for facilitating the secondary use of all types of data (including genomic data)

The current focus of radiological research on artificial intelligence (AI) has highlighted the need for large datasets including imaging data alongside detailed clinical data. Collecting such large datasets currently presents such a substantial challenge that it is only feasible in the context of EU-funded projects such as CHAIMELEON[10] and EUCAIM[11] (which also represent more detailed use-cases themselves) or by obtaining data from private, data-holding entities. Facilitating access to large datasets – as potentially included in the EHDS – would not only be economically beneficial as it would allow for easier development, validation, and monitoring of AI solutions in radiology but would also benefit patients through higher-quality research and updating of imaging biomarkers in shorter timeframes.

But even outside AI applications radiological images alone in most cases have only limited value if not associated with meaningful clinical data (including genomics e.g., in the case of central nervous system tumors).

---

[7] Montoy, J. C., Dow, W. H., & Kaplan, B. (**2016**). **Patient choice in opt-in, active choice, and opt-out HIV screening: randomized clinical trial**. *BMJ*, h6895. https://doi.org/10.1136/bmj.h6895 https://www.bmj.com/content/352/bmj.h6895

[8] Case study provided by ESR

[9] Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data (https://www.nature.com/articles/s41598-022-19045-3) / Forensic personal identification utilizing part-to-part comparison of CT-derived 3D lumbar models (https://www.sciencedirect.com/science/article/pii/S0379073818304079) / Medical Imaging and Privacy in the Era of Artificial Intelligence: Myth, Fallacy, and the Future (https://www.sciencedirect.com/science/article/abs/pii/S1546144020303859)

[10] https://chaimeleon.eu

[11] https://cancerimage.eu

In a broader context it is also important to note that much of what is technically possible to improve patient outcomes through imaging AI cannot be translated into clinical practice as access to sufficient data that would enable validation for safe clinical use is lacking. Similarly, in radiology small-scale pre-clinical research is generally favored over larger real-world research just because sufficient image data is so difficult to collect, which leads to frustration among researchers, clinicians (and companies) and potentially puts Europe at a disadvantage compared to others.

Need for inclusive data sets and the risk of exclusion of certain patient groups or data bias

Especially in the context of AI it is imperative to include diverse datasets (including minorities and edge cases) to ensure comparable performance across different patient characteristics. It has recently been shown that e.g., AI models for detection of pathology on chest X-rays show varying performance depending on patients' reported ethnicity[12]. Similar effects can be assumed for several scenarios, even in instances where potentially large datasets can be gathered in a structured way (for example in screening like breast, lung or potentially prostate). It is known that e.g., there is a correlation between ethnicity, age, and breast density / smoking history but also that certain ethnicities are less likely to join such screening programs. Opt-in solutions could therefore further increase the bias in such datasets as these ethnicities would probably be less likely to agree to data sharing even outside screening programs. To leverage healthcare data to improve outcomes for underrepresented populations and historically disadvantaged groups, it would therefore be imperative to ensure access to as much data as possible in a way where different aspects (socioeconomic status, ethnicity, genetic data *and* imaging data) can be analyzed in a comprehensive way on large datasets. Similarly, rare diseases (like sarcomas or storage diseases) lack representation in imaging-based research (even outside AI applications), hence would comparably benefit from an opt-out based approach of sharing pseudonymized data.

Need for comprehensive datasets on environmental factors and health

Environmental factors may play an important role in the development of many pathologies and their corresponding appearance in imaging. Most obviously, it can be assumed that environmental factors together with patient characteristics play an important role in lung diseases and might even influence tumor appearance e.g., on chest-CT. Access to large datasets including environmental factors, patient characteristics and imaging would enable researchers to unlock new relationships between variables and compare imaging appearances across countries or within countries but dependent on local factors like air pollution. To differentiate environmental factors sufficiently accurately from other potentially confounding variables massive datasets will be needed, but ultimately will allow to further optimize patient health.

Rheumatology[13]

Commented [MM4]: Also quite long but I think it is a nice case study

Tuberculosis risk with biologics

TNF inhibitors (drug that suppresses the physiologic response to tumor necrosis factor) entered the market after all requirements (clinical trials in place) by the year 2000. It was the first time very

[12] Algorithmic encoding of protected characteristics in chest X-ray disease detection models (https://www.sciencedirect.com/science/article/pii/S2352396423000324)
[13] Case study provided by EULAR

expensive drugs were used for a rheumatic condition, in this case, rheumatoid arthritis. For that reason, and for the type of medication they were, biologics, with long-term unknown risks, many national societies started registries.

Registries collect standardised data, directly from clinical records, and their protocols undergo all good clinical practice formalities, including consent to become participants. However, sometimes new questions arise once we start analysing the data of registries and it might not make sense, or it takes too long, to start a full project again, with protocols and new informed consent if the relevant data is already collected in clinical records despite not being part of the initial data collection forms.

When the first data was downloaded in Spain, researchers soon realised that tuberculosis was a problem, but additional information that was not part of the standardised data collection was needed. This was information for the description and evolution of the cases and potential use of preventive measures, in other words, secondary data uses. The Ethics Committee approved the amendment analysis, but no additional consent form was added.

Without having access to pseudo-anonymised data, it would not have been possiblle to identify and describe, with clinically meaningful detail, the risk of tuberculosis posed by biological medications fast enough to develop counteract measures.

More than 4 million Europeans with different forms of arthritis take these medications safely thanks to that information being available and used.

The EULAR RMD registry

People with rheumatic and musculoskeletal diseases (RMDs) in Europe are developing a registry to understand things that happen to patients that are not usually studied in medical studies and that matter to them, like interactions with outcome and social deprivation, or work or family-related issues. The idea is to create a fully patient-driven registry.

The registry will also be useful for the patients to sign up as volunteers in clinical trials for rare diseases or longitudinal observational studies for specific diseases. As many countries possible where very little information on RMDs is available due to a reduced research or rheumatology workforce are attempted to be included.

Many problems in setting up this registry were encountered due to the interpretation of the legislation on health data:

- For participants to sign-up for additional studies, ensuring an accurate diagnosis is critical. Organisers of the registry cannot double-check easily if somebody saying he or she has a disease, actually has it, as we cannot access medical data. The patient cannot either.
- It is very difficult to collect clinical data that are reported by a layperson, even if well-trained (e.g., analysis, or genetics)
- It is impossible to include any data to environmental data for assessing the risk of complications or diseases

Many relevant questions, that really matter to people with RMDs, will not be answerable.