Biomedical Data for Artificial Intelligence

Magnus Boman[1,2]
Erik Arner[4]
Carsten O. Daub[2,3]
Kazuhiro Sakurada[5]

[1]KTH Royal Institute of Technology, Stockholm, Sweden
[2]Karolinska Institutet, Stockholm, Sweden
[3]SciLifeLab, Stockholm, Sweden
[4]RIKEN IMS, Yokohama, Japan
[5]RIKEN MIH, Tokyo, Japan

Authors' contributions:
All authors participated in the symposium and in the post-symposium editing process of co-created material. This included contributing text to this paper.

The authors have no competing interests to declare.

Abstract

An overview of the state-of-the-art on biomedical data for artificial intelligence is provided via a co-created opinion statement on trends, challenges, and opportunities. The main focus is on biomedical research but the perspectives are multidisciplinary, with a special eye on how to best manage biomedical data so that it can be used for data-driven reasoning. The clinically actionable insights that could be provided by such pipelines are critically reviewed, with ethical aspects highlighted. The take home message is that while reference datasets are a potentially powerful means to new findings also of clinical relevance, the requirement specification for how to build and maintain a dataset infrastructure is difficult to perfect and requires a dedicated and goal-oriented effort from participating organisations.

1 Introduction

Almost 100 experts on biomedicine or artificial intelligence gathered at the RIKEN Center for Integrative Medical Sciences in Yokohama for two days in November of 2019. The occasion was the *6th RIKEN-KI-SciLifeLab Symposium: Biomedical Data for Artificial Intelligence* (Figure 1). As organisers, we sought to produce a tangible outcome in the form of a useful class of reference datasets from various strategic areas, including genetics/genomics, transcriptomics, epigenomics, proteomics, metabolomics, immunology, phenotyping, epidemiology, microbiome, and behaviour analysis. The main purpose of this effort was to attract AI researchers to problems in the life sciences, by giving them access to structured data. Candidate datasets were to meet criteria of:

- already being published, in order to prevent a long waiting time before they could be coordinated for use by the participating Data Centres,
- a large enough size, measured e.g. by number of samples or dimensions, in order to be suitable for data-driven processing for AI method application,
- having a well described biological background,
- coming with background knowledge of the underlying biological processes, so that test results can be evaluated and validated.

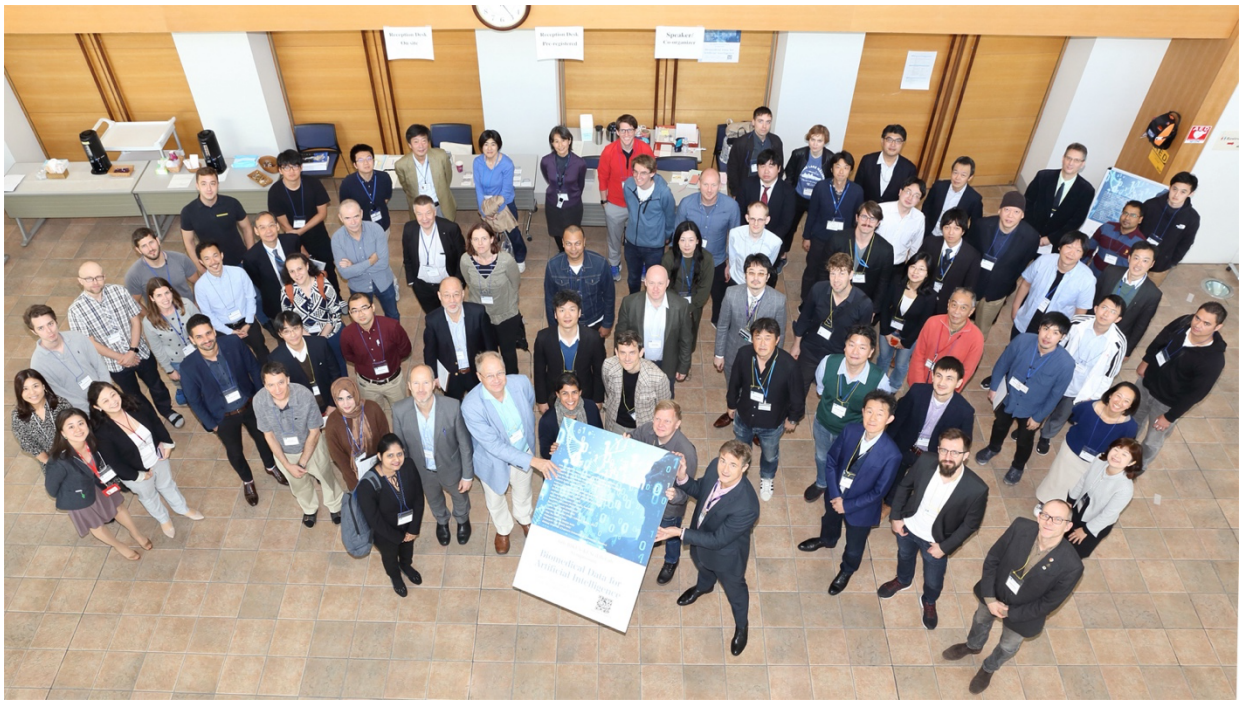

Figure 1: Most of the participants and all of the organisers, captured outside the lecture hall at IMS RIKEN, Yokohama.

At the symposium, all speakers were asked in advance to comment on some questions prepared by the arrangers of the symposium. The AI-oriented speakers were asked to prepare a ten-minute presentation of a successful application of AI or machine learning to a biomedical problem that included:

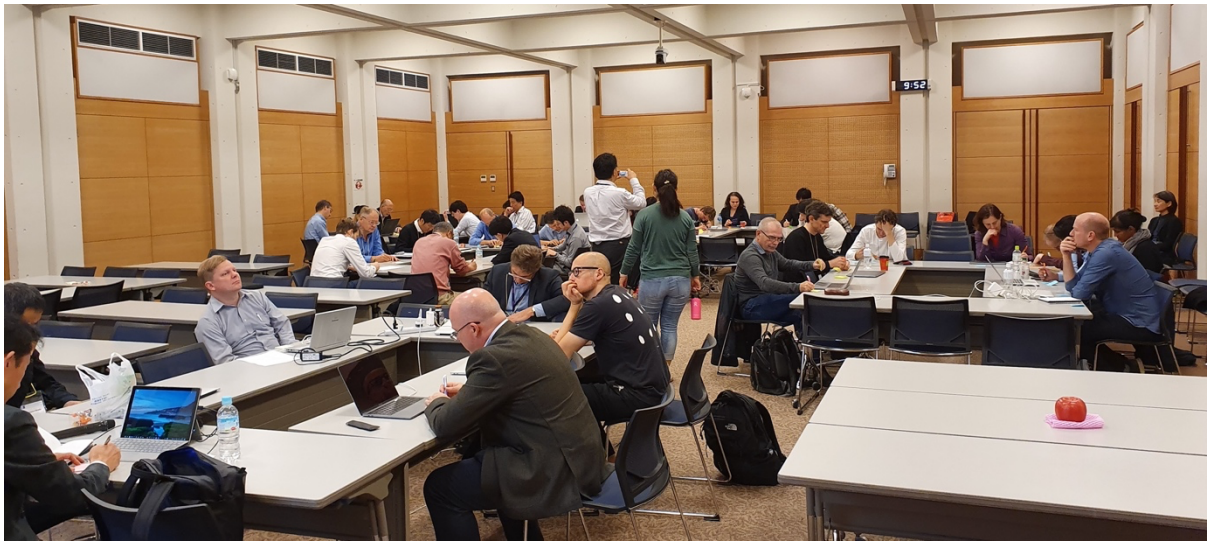

Figure 2: Participants in pensive mode for the individual idea and concept generation phase of the speedwriting workshop.

A1. Brief introduction of the biomedical problem
A2. Description of the available data, highlighting particular challenges
A3. Choice of method and why
A4. Result and wider implications

Similarly, the biomedical researchers were asked to include:

B1. Brief introduction of their research field
B2. Central question or problem that may be addressable by AI methods
B3. Any wish or expectation for what an AI approach may be able to do, potential in terms of resources saved, improvement in public health, etc.
B4. Description of the data that is typically generated or worked with that may be a candidate for a reference data set

To AI researchers and practitioners, the complicated relationships and background to data-driven investigations into biology often prove challenging. There is usually no lack of variables to measure, but complete and pre-processed datasets are hard to come by. There is also a talent problem: many labs seek to appoint lab members with backgrounds in both the medical sciences and in computational science or math. The multimodal nature of data, like when combining modalities (e.g. image-to-gene expression) can sometimes provide a shortcut to semantics (Webb, 2018). Generative models can help create pseudo-populations and synthetic data, which can in turn be refined by adversarial methods. Companies with access to large data can make incremental improvements and also occasionally provide disruptive leaps. Some of the suggestions made of AI-supported investigations were:

- Single cell networks from images
- Mechanisms behind cell heterogeneity
- Epigenetic marks as prognostic markers
- Screening for therapeutic approach and integrate host-microbiome co-metabolism
- Bridges between structured and unstructured data in biomedical domains
- Finding viral genome integration sites

Skeleton proposals for 13 different datasets were discussed, providing an excellent basis for bootstrapping future joint work. The process of grounding a roadmap towards a requirements specification was hence successful, but for the reference datasets to be put into the hands of data scientists and AI experts, ownership of the future process must be assumed by individuals, as appointed within the organisations involved. We would argue that this problem is generic, and not resulting from any structure particular to the three organisations at hand. Representatives with experience and responsibility for infrastructure and maintenance of data were present, keeping the proposals within practically feasible limits.

2 Materials and Methods

The discussion on the future of AI, data science, and data-driven methods for biomedical data opened with brainstorming, in which individually generated ideas were collected in the form of notes on trends, opportunities, and challenges (Figure 2). These notes were clustered into six groups:

- Gaps
- Mechanisms
- Ideals and Risks
- Transparency
- Augmentation
- Openness

The notes were distributed to the respective groups to support their discussions. This helped to quickly establish a focus and discussions were intense. With one person named as the writer, groups produced almost a full page each of text in only 13 minutes, using the speedwriting methodology (Boman 2014). The outcome was later refined in a continued co-creation process via a shared digital document.

3 Results

The following six subsections contain material co-created in the speedwriting groups. The actual notes that were produced in a first step of silent private contemplation and note writing are available as high resolution images in the supplementary material to this essay (repo: http://tiny.cc/RIKEN-KI-SciLifeLab).

3.1 Gaps

AI was first presented to the medical community as second opinion, an expert system that can these days be consulted via a subscription/consultation model. Typically, the predictions generated by machine learning are not contributing to causal explanations and the integration into regular care has been very slow, with AI today often being considered as first opinion instead. In application areas like digital pathology, for instance, the machine may start and human experts follow. For triaging, we may imagine walking into a hospital, and a computer decides which tests to run. On the care provider side, there is also the concept of augmented human, in which a doctor is endowed with even more expertise, thanks to informed decision support from an AI system.

A gap that may in such situations occur is the *trust gap*: Is AI curation as reliable as manual curation? Can an AI system choose which samples need human expert curators? A related gap is the *control gap*: How do humans stay in control of ethics, making sure that human interests are always served by AI systems? Old mechanisms like ethical permits do not suffice here. A *data gap* makes pre-processing necessary before AI systems can give any advice whatsoever, for example via the train-test-validation loop often adapted in machine learning. Data integration from different medical-related modalities is often necessary, like looking at genetic data and EHRs simultaneously for a patient, and combining the result in an intelligent way.

A *creative gap* points to the tendency for AI systems to do 'more of the same.' That is, instead of innovating, they play it safe by assuming that the immediate future will resemble, or even be identical, to the past. A *method gap* lies between statistical learning theory, fairly well-known to humans, and modern machine learning algorithms and methods. While the former can be quite technical and difficult to grasp, the latter can be completely opaque. A trend in AI in recent years is to abstain from optimal solutions to keep transparency and intelligibility as system properties. More pragmatic aspects of this gap is related to fixing a level of accuracy and reliability before moving an AI model into the clinic. This is related to the *explanation gap*: if an AI system helps humans in recognising patterns of a size and complexity beyond the ability of humans, how can we understand how the hidden pattern was found, and what its significance is?

The *granularity gap* lies between what we can see and what we need for successful care. Healthcare data is noisy. How much data is enough to really understand health, through the noise? It seems neither genome nor transcriptome is enough, do we need to know what is happening in each cell, or in each human? Finally, the *regulation gap*: synthetic data sets help us understand what our methods actually can tell us. What are the risks associated with having obscure computer resources in obscure places that generate fake data to confuse humans and confuse other AI systems, as in adversarial method development? Data terrorism and pollution of public datasets seem to have created a need to police, monitor, and regulate data, but how?

3.2 Mechanisms

There are at least five families of questions pertaining to mechanisms:

- Interpretability of AI-models
- Biological mechanisms that are of interest to identify and act upon
- Classical statistical learning vs machine learning
- Automation
- Scalability

A naive first step in interdisciplinary collaboration can be to make as much data as possible available to data scientists. Scalability and generalisability remain hurdles for bringing research into the clinic. We need to be extremely aware of bias, or adapt to a necessary evil. Machine learning models are notoriously hard to interpret. Furthermore, there are adversarial attacks, 1-pixel attacks, and even attacks on software intended to explicate AI results (Slack, 2018). Do we need a traditional statistical model that can be tested in a hypothesis-driven framework in order to say things about mechanisms, or can we be data-driven? Should machine learning solely be used to automate manual labour? What are the economic and social processes that affect scalable development of AI? The problem of bringing together biological and epidemiological causation is hard and we must note the difference between identifying causal

chains and manipulating causes. Both functional biology and epidemiology is about explanations and reliability and trust in models is necessary to move on to useful predictions.

3.3 Ideals and risks

What does a perfect future hold, in which biology and AI are fully integrated into regular care? And what are the associated risks? Two bullet lists concisely provide perspectives.

IDEALS

- Diagnosis and treatments will be personalised and made more effective with AI. Personal monitoring of health indicators generates data that may be used to prevent disease before it emerges, and targets are open systems, meaning new data can change the prediction target at any time as information is refined.
- Humans get an extended lifespan, long healthy aging.
- Less medication, and better targeted. Not everything needs to be treated with medicine, and learning AI methods figure comorbidities out, design smart pills, and address any required lifestyle changes. Prevention of disease, via targeted smart therapies.
- Computer-aided diagnosis is implemented everywhere it is useful.
- Health-motivated environment monitoring of the patient ecosystem and everything that surrounds an individual, in societal as well as natural habitat.
- Data available for all kinds of health purposes, including ‘blue sky research’.

RISKS

- Ageing populations and overpopulation leading to further diminished resources.
- Privacy risks with large health databases and monitoring.
- Low accuracy levels in AI predictions.
- Loss of jobs.
- Inequality risks related to AI system accessibility being unfair.
- Reliable data lacking for life science research.
- Increased monitoring and data collection to evaluate our state of health yields new risks for attacks, as such information is valuable.
- Health insurance access inequalities.
- Increased state surveillance of individuals.

3.4 Transparency

Through philosophical analyses of responsibility and ethics of autonomous robots, as well as the policy processes that drive their development, a wide debate involving priests, physicists, astronomers, politicians, and many more professions has now been going on for at least three decades (DeLanda 1991). For medical decisions, clinical guidelines and inspecting bodies are operating in our hospitals, so that whenever human expertise is questioned, there is some process to fall back on. For a new kind of decision maker, a first opinion AI system suggesting clinical action, existing guidelines and protocols can not simply be copy-pasted in. People perceive advice from a machine differently from that of a human doctor, and what if it is a panel of machines deciding? In most cases, there is not one correct diagnosis. This complicates validation of machine advice, just as human advice has changed over the last 125 years or so of modern medicine. Research turns human biology into a moving target, which will affect learning machines too.

Hybrid systems consisting of AI and humans working together to solve problems are sometimes appropriate, such as when an expert system is devised for clinical problems. Then

human experience can to some extent be coded into actionable insight, and heuristics can be passed on from human to machine, explicitly or implicitly. But not all efforts for human care are evidence-based and fit to put in a database. The systemic and holistic lessons learned for how to interpret evidence and provide humans with advice must be taken into account. Machines operating on quantitative evidence and providing advice based on that foundation are fundamentally and technologically distinct from machines operating on qualitative evidence and in an environment partly dictated by judicial aspects and norms.

For advice to be clear, the entity providing it should be transparent. In recent years, some machine learning algorithms have abstained from optimality in favour of intelligibility and transparency. Could a machine learning system even train to become more intelligible? Interpretation of computer output has traditionally been a human task, but could the two meet half-way?

The hardest problems, usually involving causal explanations and reasoning under harsh time constraints require human cooperation. AI systems have been called in to help many times over the last 20 years with various tasks related to extreme events and health-related puzzles. How to sustain collaborations between humans and machines in large, perhaps global, networks is very different from creating *ad hoc* collaborations for short-term solutions. Could machines benefit from repositories and reference data sets? What is the role of human trust and cooperation in striking up long-term collaboration for the hardest problems, with AI systems in there for decision support?

3.5 Augmentation

Even if historians often say that there is nothing particularly revolutionizing about the times we live in, healthcare seems to be in transition. Predict for good health, old age, the importance of mutations; the list is long. With more data, faster methods to analyse the data, personalised results, and reliable and fast conclusions, it is hard not to think of such progress as coming up with solutions at a much faster pace that before. This is leveraged by a model zoo, searchable model repositories, and commercially driven by companies that would love to monitor your regular health check-ups. It all leads to one future: the augmented human.

For augmentation, machines have to access profiles from individuals. How does one collect such data without compromising personal integrity? Federated systems for machine learning use distributed learning principles that can operate without having a centralised data store. Automated processing and analysis of health records is done in many countries today, while in other countries data seems to be lost for research purposes if it ends up in a health record. Biochemical profiles from our blood, monitoring of circulating tumor cells, supporting the understanding of incidence and recurrence of cancer: if we do not allow for monitoring because it may disrupt and interfere with people's lives, we may allow devices to be monitoring us ubiquitously and transparently. What transparency means rapidly becomes more opaque with the advent of qubits and quantum interference devices, and in nanotechnology, partly due to the different notions of 'observation' in these different realms.

Since the augmented human may suffer from information overload, communication with the AI part needs to be smart, especially since we are dealing with life course data. With time, logs will be long, and non-anonymous data will leak. How can such data be found and deleted, as necessary? Any massive data project requires technological security measures and here we are considering one project per individual on the planet.

3.6 Openness

In practice, the accuracy of real world data depends on the data itself. Missing values, typos, artifacts, erroneously imputed data points are all on the list of issues. If too much streamlining of health data collection and reporting are put in place, there is a risk that this unnecessarily enforces a uniform style of living, in accordance with data. That said, openness could also mean tolerance to diversity. The training of AI systems must also reflect diversity. As humans have *a priori* hypotheses and do not see the biases they build into the AI models, the models sometimes make the wrong predictions. If the sample size is small compared to the number of features, how do we train and use the AI models properly?

Openness does not automatically produce data accessibility. Data is often in isolated islands, making it hard to see the big picture. If data were only accessible to pharma, then data would chiefly be used for drug development. If data is open, we can use it for many purposes, and some of those we can not predict now. A security risk is the incentive to generate realistic fake data. Such fake data must be possible to identify and avoid. Health data is also special, as compared to for instance data from the process industry, and often even more sensitive.

Experimental data and real world data have different meanings. The former will be replicable, while in the real world, reproducibility makes no sense, as everyone is different. Experimental data is likely to fabricate the explanation based on the hypothesis. In real world data, the danger is instead that we can not analyse the full data, so we can not know if it applies to the general population. Openness thus has a role to play also in generalisability: one way of doing successful transfer learning is by training on the population of one country and testing it on that of another.

4. Discussion

Reference datasets provide AI researchers with a concrete opportunity to apply and systematically evaluate a range of methods in the context of life sciences data. Life science researchers, on the other hand, gain a deeper understanding of how their data need to be described (annotated) and structured (data format) so that they are most useful for AI research, and for which questions AI would be the method of choice for gaining biological or medical insights.
Research data is usually produced with a specific question in mind. The experimental design and employed protocols are optimized for the question at hand, which includes the number and types of sample replicates and amount of corresponding data, specifically in the context of sequencing data. Variations in biological datasets typically originate from several sources including the biological question addressed by the study design. Moreover, additional variances will be present in the data as the result from unknown sources based on experimental techniques or practical aspects.

Many clinical and basic (pre-clinical) research groups are collecting material including human samples, material from cell culturing as well as samples from animal models. Prerequisites for comparing datasets generated by different research groups include the ability to understand:

- the details about samples and conditions including experimental details, cell types and animal models, patient characteristics, etc.
- the technical details regarding the technologies used for material preparation, and experimental protocols.

Documentation should be catered to the needs of AI researchers and contain the details needed to use the dataset. The size of the datasets has to be sufficiently large to allow for the testing of AI methods. A ground truth has to be provided to allow application and supervised methods as well as for the evaluation of classification results. The access to the datasets has to be easy and the dataset has to be structured in a way that allows directly using it without further reformatting. Life sciences data are typically uploaded to a data repository such as the Sequence Read Archive (SRA) or the Gene Expression Omnibus (GEO). Such repositories require describing the provided dataset with a rather minimal level of detail limiting the usability of such datasets beyond the purposes of the original publication. Also, aspects important for analysis of the dataset with AI methods are not considered by such repositories. Additional data that may also capture aspects relevant for further analysis are often published as supplemental tables, not adhering to any specific format and often without explicit mappings to the repository data. Data Centres facilitate the description and collection of research data produced at universities or research institutions. The main goals of Data Centres include making research data reusable beyond the initial aims for which these data have been produced.

Explanation is a central goal of science. At the same time, most prediction procedures developed using machine learning and multimodal clinical data are essentially separate from causal explanation. We would argue that explanation and prediction are best understood in light of each other, which presents barriers, including:

- The quality of human data is an obstacle to accurate predictions.
- To correctly describe time series data in computable format that ties to a causality model is an obstacle to apply AI in data analysis.
- There is still a need to develop quantitative means to identify distance and similarity between different sommatic and mental states represented by multimodal variables, and to predict such states for each individual.
- A frame of guidance for developing an assured and/or trustworthy architecture for medical data-driven solutions in health care.

In our particular context, the question of ownership of the process for developing and maintaining reference datasets have been presented to the leaders of the three participating organisations. We note that there are low-hanging fruits in that some reference datasets could be named and described in plenum at the symposium, but also that a sustainable plan needs to be developed for how to best make datasets available, and how to feedback results and opinions to their owners. This is a generic problem not related to the three organisations as such. A year after the symposium, an online two-hour mini-symposium were held, at which the preliminary contents of this essay were presented. The main stakeholders from the three institutions were present: the leadership, principal investigators, database experts, reference data set owners and their prospective future users. This closed the loop of the need for reference data sets voiced at the fifth symposium in the series, and discussed at length at the sixth. This essay will be shared as a firestarter for discussions at the seventh, planned for Stockholm in the autumn 2021.

Acknowledgements

The authors would like to thank all the participants of the Yokohama symposium for generously sharing their thoughts in the speedwriting session. Sabine Koch provided important suggestions for improvement on an earlier version of this paper.

References

1. Boman, M. 2014. Speedwriting in networked foresight, in *Innovation for Sustainable Economy & Society: The Proceedings of The XXV ISPIM Conference*, The International Society for Professional Innovation Management.
2. DeLanda, M. 1991. *War in the Age of Intelligent Machines*. Zone Books.
3. Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. 2019. How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods. *arXiv preprint arXiv:1911.02508*.
4. Webb, S. 2018. Deep learning for biology. *Nature*, *554*(7693). DOI: 10.1038/d41586-018-02174-z