A good heart for ethical AI

Exit report for Ahus sandbox project (EKG AI) Theme: Algorithmic bias and fair algorithms

February 2023

Datatilsynet

Contents

What is the sandbox?

In the sandbox, participants and the Norwegian Data Protection Authority jointly explore issues relating to the protection of personal data in order to help ensure the service or product in question complies with the regulations and effectively safeguards individuals' data privacy.

The Norwegian Data Protection Authority offers guidance in dialogue with the participants. The conclusions drawn from the projects do not constitute binding decisions or prior approval. Participants are at liberty to decide whether to follow the advice they are given.

The sandbox is a useful method for exploring issues where there are few legal precedents, and we hope the conclusions and assessments in this report can be of assistance for others addressing similar issues.

Note

Technology and the law are constantly evolving, and adjustments and clarifications may have occurred after this report was written.

Summary

The goal of this sandbox project has been to explore the concepts of "fairness" and "algorithmic bias" in a specific health project, EKG AI. Akershus University Hospital (Ahus) is developing an algorithm for predicting the risk of heart failure in patients. In time, it will be used as a decision-support tool[1] to enable health personnel to provide better and more effective treatment and follow-up of patients. In this sandbox project, we have discussed the possibility of bias in EKG AI, as well as potential measures to prevent discrimination.

Summary of results:

- **What is fairness?** The concept of "fairness" has no legal definition in the General Data Protection Regulation (GDPR), but it is a central principle of privacy according to Article 5 of the Regulation. The fairness principle is also central in other legislation, and we have looked to the Norwegian Equality and Anti-Discrimination Act to clarify what the principle entails. In this project, we have assessed EKG AI's level of fairness with respect to non-discrimination and transparency, the expectations of the data subject and ethical considerations of what society considers fair.

- **How to identify algorithmic bias?** To ensure the algorithm is fair, we have to find out if the EKG AI algorithm returns less accurate predictions for some patient groups. In this project, we chose to look closer at discrimination on grounds of "gender" and "ethnicity". When checking the algorithm for discrimination, one would normally need to process new personal data, including special categories of personal data. In this context, one must consider the requirements for the legality of processing and the requirements of the principle of data minimisation for proportionate and necessary processing of personal data.

- **Which measures could reduce algorithmic bias?** This sandbox project has highlighted a potential risk of the EKG AI algorithm discriminating against some patient groups. Bias can be reduced through technical or organisational measures. Potential measures for EKG AI include ensuring that the data source is representative, and making sure health personnel have good information and training to make sure the predictions are applied correctly in practice. In addition, Ahus will establish a mechanism for monitoring the accuracy of the algorithm and making sure the algorithm is trained as needed.

Going forward

Ahus wants to try out the algorithm in a clinical setting from early 2024. Clinical decision-support tools based on artificial intelligence (AI) are considered medical technical equipment and require a CE marking issued by the Norwegian Medicines Control Authority in order to be implemented in clinical activity.

This sandbox project has highlighted a potential risk of the EKG AI discriminating against some patient groups. Ahus will consider the possibility of conducting a clinical trial to explore whether the algorithm is less accurate and produces less accurate predictions for patients with different ethnic backgrounds (in this report, ethnic background refers to genetic origin). The results of the trial will indicate whether corrective action is necessary in the algorithm's post-training phase.

In the project period, we have discovered that there is no common, baseline method for identifying algorithmic bias. If we had had more time in the project, we would have developed our own method, based on experiences gained in the project period. In addition, it would have been interesting to dive even deeper into the ethical requirements for the use of artificial intelligence in the health sector.

---

[1] Decision-support tools = The preparatory works to the Health Personnel Act make it clear that the term "decision-support tool" shall be broadly understood and that it encompasses all types of knowledge-based aids and support systems, which may provide advice and support, and may guide healthcare personnel in the provision of medical assistance.

About the EKG AI project

Akershus University Hospital (Ahus) is a local and regional hospital, with 12,000 employees. Ahus is responsible for a population of approx. 594,000 in the Follo, Romerike and Kongsvinger region, as well as the northernmost districts in Oslo. Ahus is Norway's largest emergency hospital, with patient services covering somatics, mental health care and addiction treatment.

Ahus has developed a decision-support tool based on artificial intelligence, EKG AI, which can predict heart failure in patients. The decision-support tool has been developed by linking EKG data to specific diagnoses, so-called supervised learning[2]. After training, testing and validation, this process produces an algorithm capable of predicting the probability of heart failure. No similar tools have previously been implemented in clinical activity. With Ahus's large patient base, the project is in a good position to develop a tool with high level of accuracy, that can also be implemented in other hospital trusts in Norway.

With EKG AI, Ahus seeks to:

- Increase efficiency in diagnosing and treating heart failure.
- Improve the diagnostic process for heart failure and enable heart failure to be determined sooner.
- Reduce the time patients have to stay in hospital, treatment times and mortality rates.
- Commercialise the algorithm in the wider health sector.

Data sources and dataflow

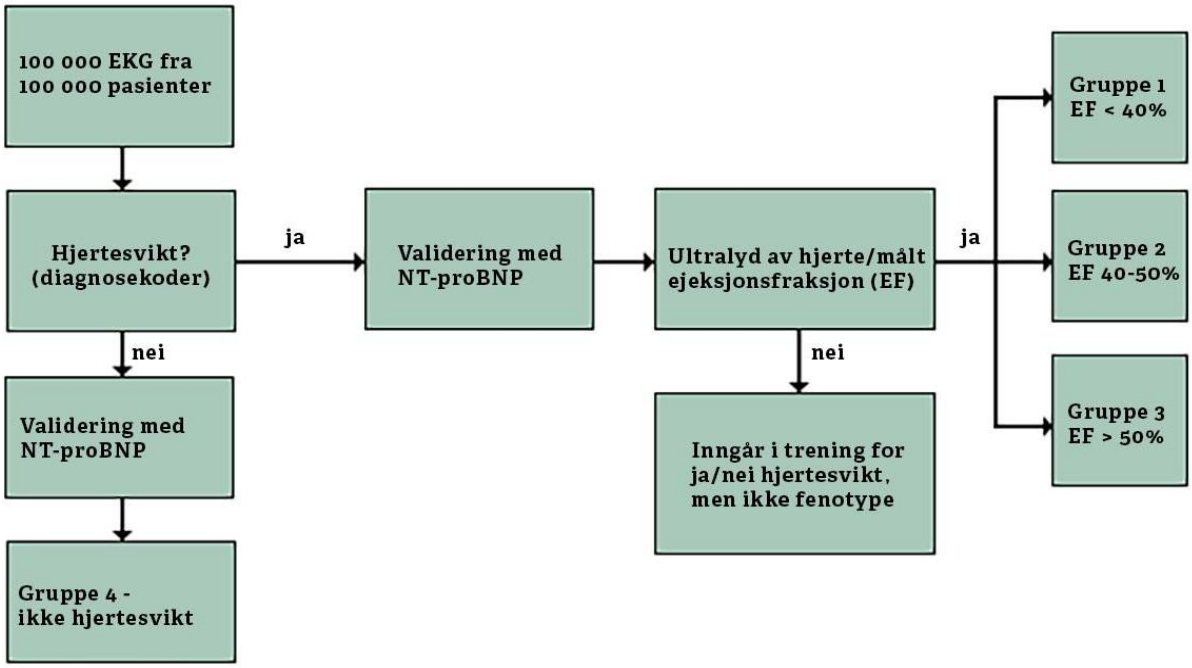
The decision-support tool is being developed on a Google Cloud platform with the autoML tool Vertex AI.

The system was fed approx. 100,000 EKG scans from patients who have been admitted to Ahus in recent years. There are three different phenotypes of heart failure, each requiring different approaches to treatment. The EKG scans are grouped accordingly:

- Group 1: heart failure with reduced ejection fraction
- Group 2: heart failure with mid-range ejection fraction
- Group 3: heart failure with preserved ejection fraction
- Group 4: no heart failure

The grouping is based on diagnostic codes, NT-proBNP blood tests and ultrasound images of ejection fraction (EF). See figure on the next page.

[2] Supervised learning means categorised data is used. The supervision is the labelling that accompanies the data. From Artificial Intelligence and privacy, Norwegian Data Protection Authority, 2017, p. 7



The data is taken from two different sources: The EKG archive ComPACS (EKG readings and ejection fraction (EF) and the patient file system DIPS (diagnostic codes and NT-proBNP blood tests). After being categorised into four groups, all EKG readings are pseudonymised, before they are transferred to the Google Cloud platform for training, testing and validation with Vertex AI.

> **Pseudonymisation**
>
> De-identifying personal data so that they cannot be linked to a specific person without the use of additional information (such as a key), which is stored separately under adequate protection. Pseudonymised personal data is not anonymous.

Objective of the sandbox project

Ahus had already started developing the EKG AI algorithm when it was selected to participate in the Data Protection Authority's regulatory sandbox in the spring of 2022. Ahus wanted to discuss algorithmic bias and how to ensure that the EKG AI returns fair predictions.

There is limited judicial precedent on the requirement of fair algorithms, and the GDPR does not give any clear answers as to how this principle should be interpreted in practice. The objective of this sandbox project has therefore been to explore what a principle of fairness, as established in Article 5, entails, and how this should be interpreted in the context of a specific AI project.

The goal of the sandbox project has been to explore whether the EKG AI has bias, and to propose specific measures to reduce any such bias. The intention behind the measures is to develop algorithms that promote equal treatment and prevent discrimination. In this context, the Equality and Anti-Discrimination Ombud (LDO) has contributed valuable expertise on discrimination legislation.

The purpose of this exit report is to communicate the discussions and results of the project to a wider audience, as these may have transfer value for other health projects that use artificial intelligence.

Objectives of the sandbox project:

1. **What are fairness and algorithmic bias?** Gain a better understanding of the concepts "fairness", "algorithmic bias" and "discriminating algorithms", as well as to account for regulations that are relevant in this context.

2. **How to identify algorithmic bias?** Explore whether, and if so, to what degree, algorithmic bias exists, or may arise, in Ahus's EKG AI algorithm.

3. **Which measures could reduce algorithmic bias?** Propose technical and organisational measures to reduce and correct any bias in the algorithm.

Due to considerations of scope, the sandbox project has not considered legal basis requirements for the processing of personal data in this project. We briefly mention, however, that the use of patient data in the development of the decision-support tool was approved by the Norwegian Directorate of Health in January 2022 pursuant to Section 29 of the Health Personnel Act. This decision grants an exemption from confidentiality and constitutes a supplementary legal basis for the processing of personal data pursuant to the GDPR. For more information about the legal basis for processing of health information in connection with the development of algorithms in the health sector, see the exit report for the sandbox project Helse Bergen, published in late 2022[3].

Under Article 35 of the GDPR, a data protection impact assessment (DPIA) is required if the processing of personal data is likely to result in "a high risk" to the rights and freedoms of natural persons.[4] Before Ahus started developing the algorithm EKG AI, they did prepare a DPIA, but this has not been a focus of the sandbox project. However, elements from this report, particularly the method for identifying algorithmic bias, as well as the measures for reducing the risk of algorithmic bias, would be relevant to include in a DPIA.

---

[3] https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/helse-bergen-sluttrapport-kunstig-intelligens-i-oppfolging-av-sarbare-pasienter/

[4] Read more about DPIA requirements on the Data Protection Authority website: https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/vurdere- personvernkonsekvenser/vurdering-av-personvernkonsekvenser/

Fairness and algorithmic bias

In everyday speech, *fairness* is used to describe whether there is an equal and fair distribution of burdens and benefits in society. While the concept of "fairness" has no legal definition in the General Data Protection Regulation (GDPR), it is a central principle of privacy according to Article 5 of the Regulation. It provides that "personal data shall be processed lawfully, fairly and in a transparent manner in relation to the data subject".

The content of the fairness principle is dynamic, which means it changes over time, in line with society's interpretation. The European Data Protection Board (EDPB) specifies, in its guidelines on data protection by design and by default,[5] that the fairness principle includes non-discrimination, the expectations of the data subject, the wider ethical impact of the processing and respect for the data subject's rights and freedoms. In other words, the fairness principle has a wide scope of application.

A similar description can be found in the recitals[6] of the GDPR, which emphasises that this principle entails that all processing of personal data must be carried out with respect for the data subject's rights and freedoms, take into account the reasonable expectations of the data subject for what the data will be used for. Transparency in processing is therefore closely linked to the fairness principle. Adequate information is a key factor in making sure the processing is predictable for the data subject and in enabling the data subject to protect their right to fair processing of their personal data. This exit report does not discuss issues related to the information requirement, but we refer to the exit report for the Helse Bergen sandbox project[7], which does.

The recitals[8] of the GDPR further specifies that the controller shall prevent discrimination of natural persons on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation. The recitals is not legally binding, but can be used as a guide in interpreting the provisions of the GDPR. In the Ahus project, we have, among other things, considered whether the EKG AI algorithm will give all patients equal access and equally good health services, regardless of whether the patient is male or female, or whether the patient has a different ethnic background than the majority of the population.

The fairness principle is a central element in several other pieces of legislation, including various human rights provisions and the Equality and Anti-Discrimination Act. These are relevant for the interpretation of *fairness*, as their requirements are sometimes more stringent and specific than the GDPR.

What does the Equality and Anti-Discrimination Act say?

The Equality and Anti-Discrimination Act prohibits discrimination on the basis of "gender, pregnancy, leave in connection with childbirth or adoption, care responsibilities, ethnicity[9], religion, belief, disability, sexual orientation, gender identity, gender expression, age or any combinations of these factors, see Section 6.

Discrimination is defined as unfair differential treatment and may be "direct" or "indirect", see Sections 7 and 8.

Direct differential treatment means that a person protected from discrimination is treated worse than other comparable persons, see Section 7, whereas indirect differential treatment means that an apparently neutral provision, condition or practice leads to a person protected from discrimination being put in a worse position than others, see Section 8. Indirect discrimination may, for example, arise by an apparently neutral algorithm being used indiscriminately for all patient groups. As a result of the prevalence of heart failure being consistently higher among men than women,[10] female cardiac patients would be less represented in the data source, which could mean the predictions for women may be less accurate. Both direct and indirect discrimination requires a causal link between the differential treatment and the basis of discrimination, i.e. that a person is put in a worse position due to their gender, age, disability, ethnicity, etc.

---

5 Guidelines 4/2019 on Article 25 Data Protection by Design and by Default | European Data Protection Board (uropa.eu)

6 Recital 39, https://www.privacy-regulation.eu/en/recital-39-GDPR.htm

7 https://www.datatilsynet.no/regelverk-og-verktoy/sandkasse-for-kunstig-intelligens/ferdige-prosjekter-og-rapporter/helse-bergen-sluttrapport-kunstig- intelligens-i-oppfolging-av-sarbare-pasienter/

8 Recital 71, https://www.privacy-regulation.eu/en/recital-71-GDPR.htm

9 It follows from the provision that the term "ethnicity" refers to, among other things, national origin, heritage, skin colour and language.

Furthermore, public authorities have a duty to act, pursuant to Section 24 of the Act, by making active, targeted and systematic efforts to promote equality and prevent discrimination. The sandbox project may be used as an example of a project with the intention of promoting equality and preventing discrimination by implementing specific measures in the public health services.

Fairness as an ethical principle

Fairness is also an ethical principle, which means that ethical considerations are central in the interpretation, and application, of the fairness principle in the GDPR. "Ethics guidelines for trustworthy AI"[11], a report prepared by an expert group appointed by the European Commission, mentions three main principles for responsible artificial intelligence: lawful, ethical and robust artificial intelligence. The same principles are reflected in the Norwegian Government's National Strategy for Artificial Intelligence[12] from 2020.

In ethics, one considers how one *should* behave and act to minimise the ethical consequences of using artificial intelligence. While something may be within the law in the legal sense, one may still ask whether that something is ethical. Ethical reflections should be encouraged in all stages of the algorithm's life – in development, in practice, and in the continual learning stage.

In ethics, one wants to answer questions about "what is good and bad, good and evil, right or wrong, or fair and equal treatment for all"? An ethical approach to the EKG AI project would be to ask whether the algorithm will give equally good predictions for all patients. With increased use of algorithms in clinical settings in the future, it will be relevant to ask whether the general benefits to using artificial intelligence will truly benefit all patients.

Algorithmic bias

All solutions based on code will, naturally, contain errors or inaccuracies. This is true for both highly advanced systems and simpler solutions, but the more complex the code, the greater the risk of errors. In machine-learning solutions, errors will likely result in the algorithm's predictions being less accurate or incorrect. Errors that systematically produce less accurate, or incorrect, predictions for some groups will be examples of what we call algorithmic bias.

When two patients have virtually identical health needs, they should receive equally comprehensive health care services – regardless of ethnicity, gender, disability, sexual orientation, etc. However, if an algorithm recommends different levels of help, there may be reason to suspect some form of discrimination.

There are many potential reasons for algorithmic bias, and the cause of it is often complex. In this project, we have focused on four causes for algorithmic bias. This is because these four causes have been most relevant for our project, but these are also common for artificial intelligence in general. The description and illustration below are based on

Artificial intelligence

In the Norwegian Government's National Strategy for Artificial Intelligence, artificial intelligence was defined as "systems [that] act in the physical or digital dimension by perceiving their environment, processing and interpreting information and deciding the best action(s) to take to achieve the given goal. Some AI systems can adapt their behaviour by analysing how the environment is affected by their previous actions." The term *algorithm* is often used about the code in a system, i.e. the recipe for what the system is finding a solution for.

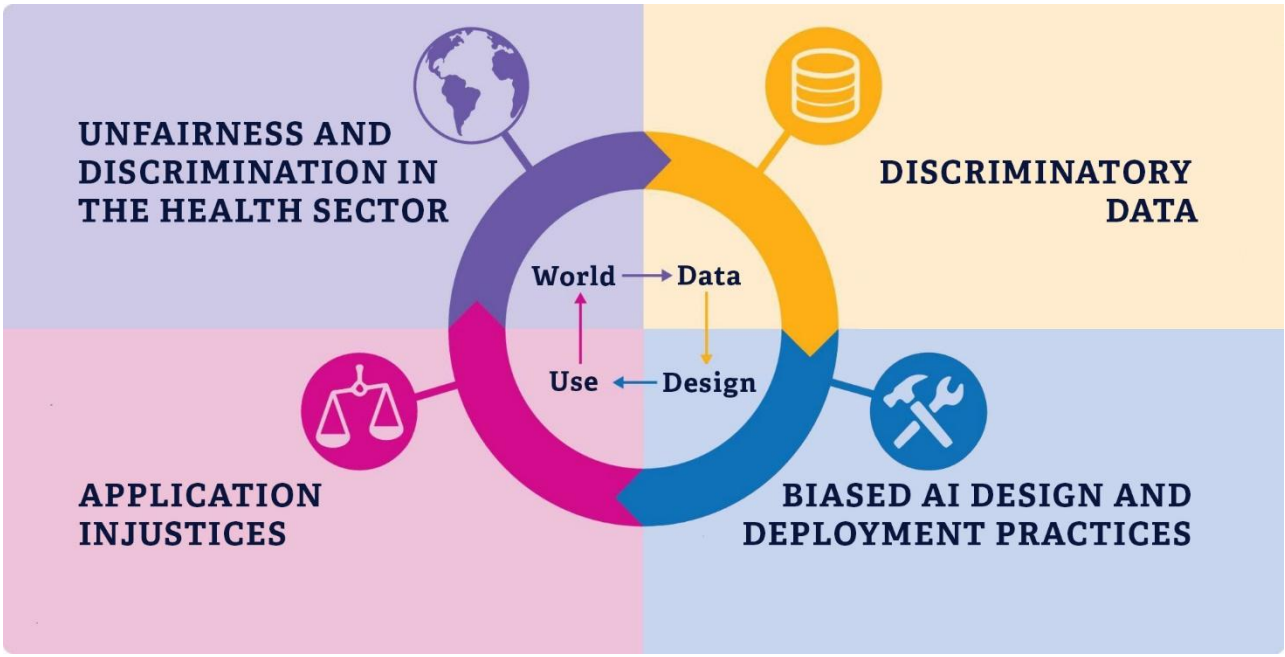
---

10 Hjerte- og karregisteret: Rapport for 2012–2016. From https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2016/hjerte--og-karregisteret.-rapport-for-2012-2016.pdf

11 Ethics Guidelines for Trustworthy AI , 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

12 National Strategy for Artificial Intelligence, 2020, <https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/>

a presentation in *the British Medical Journal*, which focuses on algorithmic bias in the health sector. We want to emphasise that several of the causes may lead to the same bias, and that there is some overlap between the different causes.



1 Unfairness and discrimination in the health sector

A machine learning algorithm makes predictions based on the statistical probability of certain outcomes. The statistics will be based on historical data, which reflects reality, including existing unfairness and discrimination in the health sector. This includes exclusionary systems, prejudiced health personnel or varying access to health services. This is perpetuated and reinforced in the algorithm.

2 Bias in the data source

Unfairness in society will be reflected in the data sources fed to the algorithm during training. If there is bias in the facts, the algorithm's predictions will reflect this bias. If the training data is not sufficiently diverse, the algorithm will not be able to give accurate predictions for under-represented individuals or groups. This could, among other things, happen when different population groups have different levels of access to health services for socioeconomic reasons.

3 Bias in design and production practices

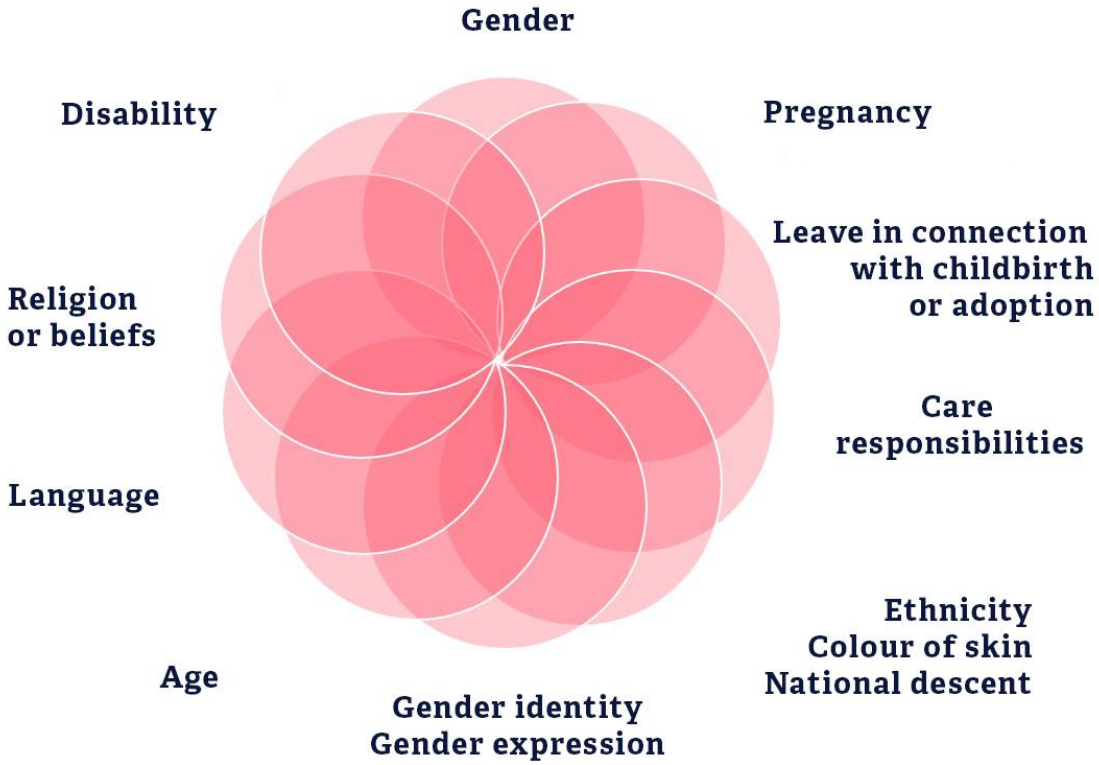
Algorithmic bias may also arise from the developers' biases and choices made in the development stage. One reason for this could be that the developers either are not aware of or do not understand the potential discriminatory outcomes of the design choices they make. Another reason could be that no systems exist to identify unintended discrimination when the algorithm is used.

4 Unfairness in application

An algorithm with hidden bias will reinforce an already discriminatory practice. In cases where the algorithm learns from its own predictions, this bias will be further reinforced in the algorithm. Using an algorithm programmed with the wrong purpose could also lead to discrimination in practice, for example when the purpose of identifying a person's health needs is justified by how much money that person has spent on health services.

How to identify algorithmic bias?

Artificial intelligence has great potential for improving diagnostics and treating disease. At the same time, there are several examples of artificial intelligence largely perpetuating and reinforcing existing social discrimination. The sandbox project wanted to explore how potential bias in the EKG AI algorithm could be identified. To ensure a fair algorithm, we wanted to see if EKG AI would give less accurate predictions for some patient groups. In consultation with the LDO, we used the bases of discrimination[13], as defined in Article 6 of the Equality and Anti-Discrimination Act, as our starting point. We learned that sometimes it is not sufficient to look at these bases of discrimination in isolation. We also need to cross-check them, such as in "minority woman".



In Norway, we have free universal health care, which is a good starting point for a representative data source with a wide variety of patients. There may still be situation, however, where certain patient groups have not had the same access to a specific treatment, which could give rise to bias in the algorithm. In this project, we chose to explore whether EKG AI's predictions varied in accuracy for patients with a different ethnic background. One of the reasons why we chose to focus on discrimination on grounds of ethnicity, was that there are several examples[14] of algorithms that systematically discriminate against ethnic minorities. From a health perspective, patients with different ethnic backgrounds can have different symptoms of heart failure, variations in EKG readings and blood tests. In this context, it is important to point out that the term "ethnicity" is used as a reference for the patient's biological or genetic origin.

The challenge for Ahus is that there is limited or no data on the patient's ethnicity. When Ahus does not have access to this information, it will not be possible to check whether the algorithm makes less accurate predictions for these patient groups. In order to determine whether such bias exists, Ahus will have to conduct a clinical trial. In such a trial, information about ethnicity may be collected based on consent, and it will be possible to see whether ethnic minorities systematically end up in a worse position than the majority of patients that the algorithm has been trained with.

---

[13] In order for differential treatment to be considered discrimination pursuant to the Equality and Anti-Discrimination Act, it must be based on one or more bases of discrimination. Read more about bases of discrimination on the Anti-Discrimination Tribunal website: https://www.diskrimineringsnemnda.no/klagegrunnlag/diskriminering

[14] E.g. an algorithm that discriminates against ethnic minorities in the health sector: https://www.wired.com/story/how-algorithm-favored-whites-over-blacks- health-care/
And Amazon's recruitment algorithm, which systematically discriminated against female applicants: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

To check for bias in EKG AI, a limit value for statistical uncertainty must be defined for the algorithm. As prejudice and discrimination will always exist in the real world, algorithms will also include errors and biases. The most important question has therefore been to determine where the limit is for unacceptable differential treatment in EKG AI.

Wanted vs. unwanted differential treatment

The main purpose of EKG AI is to prioritise patients with active heart failure over those who do not have heart failure. This is a form of differential treatment. The determining factor in an assessment pursuant to anti-discrimination legislation is whether the differential treatment is unfair.

In order to define such a limit value, one must conduct a clinical trial that explores whether the algorithm makes less accurate predictions for certain patient groups based on one or more bases of discrimination. If it turns out, for example, that EKG AI prioritises patient groups on the basis of ethnicity rather than medical factors, such as EKG, diagnosis codes, etc., this will be considered unfair differential treatment pursuant to the legislation.

When checking whether the algorithm discriminates, it will normally require the collection and processing of new personal data. A new purpose therefore requires a new assessment of the lawfulness of the processing pursuant to Article 6, and potentially Article 9, of the GDPR.

It is noteworthy that the European Commission, in Article 10 (5) of the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), would allow the processing of special categories of personal data, such as health information, to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction. If this provision is adopted, it would provide a legal basis for, in certain situations, examining whether algorithms discriminate when processing special categories of personal data.

The data minimisation principle established by Article 5 of the GDPR would still limit the types of personal data that could be processed. The principle requires that the data processed be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.[15] The requirement of necessity also includes an assessment of the proportionality of the processing. When assessing whether Ahus should collect and process information about the patients' ethnicity, we have therefore deemed it proportional in the context of the consequences a potential bias may have for the individual patient.

The predictions from EKG AI are only one of many sources of information for how health personnel assess the patient's treatment. This means that the consequences of a potential algorithmic bias would have minimal potential for harm. If the algorithm, for example, fails to detect heart failure in a patient (false negative), the heart failure could still be identified with an ultrasound of the heart and subsequent blood tests. Ahus must therefore ask whether the collection and processing of information about ethnicity, which is considered a special category of personal data, constitutes proportionate processing when the goal is to identify potential algorithmic bias.

There are no clear answers to how such a limit is to be defined. In some cases, it will only be possible to say whether the processing was necessary to identify discrimination *after* the information has already been collected and processed. In a medical context, however, it is often as important to identify what is not relevant, as what is relevant for sound medical intervention.

---

[15] Article 5 (1) (c) of the GDPR

Measures to reduce algorithmic bias

In the sandbox project, we have discussed how Ahus can reduce algorithmic bias, both in the model itself through technical measures, and in correcting for algorithmic bias after the algorithm has been implemented, through organisational measures.

It is pointed out in recital 71 of the GDPR that fair processing of personal data entails "implement[ing] technical and organisational measures appropriate to ensure, in particular, that factors which result in inaccuracies in personal data are corrected and the risk of errors is minimised, |and] secur[ing] personal data in a manner that takes account of the potential risks involved for the interests and rights of the data subject and that prevents, inter alia, discriminatory effects on natural persons (...)".

Comment from the Equality and Anti-Discrimination Ombud:

A lack of precision in the algorithm for defined patient groups does not necessarily mean that Ahus, as a legal entity, is discriminatory in its practices. The determining factor under the Equality and Anti-Discrimination Act is that persons protected from discrimination are not "treated worse than others", see Section 7. The determining factor in this context is whether the differential treatment leads to "harm or detrimental effects for the person subjected to the differential treatment, e.g. by the differential treatment causing the person to lose benefits, suffer an economic loss or have fewer options compared to others in a similar situation. There must be a specific and direct effect on specific natural persons."[16]

The Ombud is of the opinion that Ahus can compensate for algorithmic bias by supplementing with other medical methods in their examination of patient groups for which the algorithm produces less accurate predictions. The determining factor for achieving equal treatment for all is to know *which* patient groups the algorithm is less accurate for, and to *implement measures* to ensure that these patients are afforded the same level of health care as other groups.

Some general examples of technical measures include pseudonymisation, encryption or anonymisation of personal data, to minimise the intrusion on the data subject's privacy. Organisational measures include the introduction and implementation of procedures and practices that ensure compliance with laws and regulations. For more information about the implementation of suitable measures in software development, see the Data Protection Authority website.[17]

Three types of measures

In this project, we have primarily focused on three different types of measures:

1. Analysing and checking the algorithm's data sources (technical measure)
2. Establishing procedures for training health personnel in the use of the decision-support system (organisational measure)
3. Establishing a monitoring mechanism for continual learning (technical measure)

16 Prop. 81 L (2016-2017), Ch. 30, comment to Section 7

17 Data Protection Authority website on data protection by design https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/innebygd- personvern/programvareutvikling-med-innebygd-personvern/

Technical measure: Analysing and checking the algorithm's data sources

The Norwegian National Strategy for AI[18] highlights distortions in the underlying data as a particular obstacle to inclusion and equitable treatment. This is explained as follows: *"datasets used to train AI systems may contain historic distortions, be incomplete or incorrect*. Poor data quality and errors in the data source for EKG AI will therefore be embedded in and reinforced by the algorithm and may lead to incorrect and discriminatory results. Poor data quality can, for example, be caused by repeated misdiagnoses by health personnel. In order for the algorithm to be able to find an accurate pattern in the data set, the information must be consistent and correspond to facts.

Ahus has used historical health data with a statistical correlation with the risk of heart failure. The data source for the algorithm includes results from approx. 100,000 electrocardiograms (EKGs), ICD-10 diagnosis codes for heart failure and information about the heart's ability to pump blood. This information is taken from the cardiology system and the patient file system DIPS at Ahus.

In our discussion, we have focused especially on two different methods for minimising algorithmic bias in the data source:

- **Method 1** does not introduce new data to the algorithm; instead, under-represented patient groups are "inflated" in the data source. The challenge with this method is that it yields better accuracy for some patient groups, but would lead to less accurate results for patients in the majority group.

- **Method 2** adds more data points (labels) to the algorithm for the under-represented patient group. With this approach, one would have to accept a higher degree of discrimination in an initial phase, but over time, the algorithm will become more accurate. The challenge with this method is that one does not necessarily have the information needed to correct the bias that has been identified.

Only after a representation bias in the algorithm has been documented is it possible to implement corrective measures. A central focus of our discussions have been to identify which groups are under-represented in the algorithm's data source. In particular, we have looked into representation of gender and ethnicity, as there are examples[19] of other algorithms that turned out to discriminate on the basis of these factors.

Position of the Equality and Anti-Discrimination Ombud: In connection with the Ahus EKG AI project, the Ombud's position is that the greatest potential for discrimination is tied to representation bias in the data source.
Research literature[20] on heart failure points out that heart rates vary for different ethnicities. This means that ethnic minorities could have EKG curves that differ from the majority population in Norway, and this should be accounted for, both in the development and application of the algorithm.

Information about ethnicity is not recorded in the patient file, nor is this information available in any other national sources. This leaves Ahus with limited options for checking whether the algorithm is less accurate for ethnic minorities than it is for the majority population. In order for Ahus to gain access to this information, they must conduct a clinical trial[21] based on voluntary collection of data, to verify the algorithm's predictions for this patient group. In the sandbox project, Ahus has argued for a need to register information about the patients' genetic origin to ensure that they provide safe health services to all. Registration of ethnicity is a sensitive topic, and whether, and if so, how, this can be handled in practice, will be for national authorities to assess.

---

[18] National Strategy for Artificial Intelligence, 2020, Chapter 5, <https://www.regjeringen.no/no/dokumenter/nasjonal-strategi-for-kunstig-intelligens/id2685594/>
[19] Discrimination of ethnic minorities in the health sector: https://www.wired.com/story/how-algorithm-favored-whites-over-blacks-health-care/
And Amazon's recruitment algorithm, which systematically discriminated against female applicants: https://www.reuters.com/article/us-amazon-com-jobs-automation- insight-idUSKCN1MK08G
[20] Including: JCDD | Free Full-Text | The Impact of Ethnicity on Athlete ECG Interpretation: A Systematic Review | HTML (mdpi.com)
and
Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis - PMC (nih.gov)
[21] According to Helsenorge.no, a clinical trial is defined as "research on the effect of new medications or treatment methods, and on whether side effects are acceptable.", from: https://www.helsenorge.no/kliniske-studier/om/

Information about the patient's gender is part of the EKG AI data source, categorised as "man" and "woman". There is no data on whether the algorithm is less accurate for patients who do not identify as either a man or a woman, or patients who have had gender reassignment surgery. If so, a clinical trial must be conducted to check for this. The prevalence of first-time cases of heart failure has been consistently higher among men than women.[22] According to Ahus, different types of heart failure require different treatments and follow-up, and women are more represented in one of the three types of heart failure. As a result of women being historically under-represented in medical research, there is a general risk that women will have less accurate results than men when the algorithm has been trained on historical health data. Ahus can confirm, however, that there is a robust data source for female heart failure patients in EKG AI, and that it is therefore unlikely that EKG AI would discriminate against women.

If the algorithm is sold to other actors in or outside of Norway in the future, there is a risk of the data source not accurately reflecting the new patient groups the algorithm is asked to give predictions about. To maintain a high level of accuracy, the algorithm will need continual learning on new, localised patient data. This is covered in measure 3 "Monitoring mechanism for continual learning".

Organisational measure: Algorithm meets health personnel

Until now, development of the decision-support tool EKG AI has taken place in a lab. A clinical trial is necessary to verify the algorithm's accuracy and predictions on real data before it is implemented in a clinical setting. In order to optimise how the decision-support tool functions, the result of the algorithm's calculations must be presented to health personnel in a way that makes the prediction work as intended, i.e. as a decision-support tool for more efficient diagnosis of heart failure than what we currently have. The result must be available to the recipient immediately, health personnel must be able to interpret the result and apply it correctly. The overall goal is for EKG AI to be more accurate and precise than health personnel are able to be on their own, so that it contributes to improved and more efficient health services.

Article 22 of the GDPR prohibits decisions based solely on automated data processing. In order for the prohibition to apply, the decision must be fully automated, without any human involvement. EKG AI will be used as a decision-support tool, and the prohibition of Article 22 therefore does not apply. Good information and training of health personnel about how to use the decision-support tool will ensure actual human intervention is involved in the decision-making process, and will also reduce the risk of health personnel relying indiscriminately on the algorithm's predictions in practice.

Algorithms are advanced technical systems, and health personnel must know how they work and be trained in how to use them. This entails, primarily, that health personnel receive an explanation on how to use the algorithm, but also that they be informed about risk factors and margins of error in the algorithm's predictions. Health personnel must be able to understand how the algorithm works and the background for its predictions, to prevent mistrust in the tool. Insight into and an understanding of how the model works is also essential to enable health personnel to independently assess the predictions. If, in the future, it is determined that the algorithm's accuracy varies for different patient groups, this must also be communicated to health personnel. Training can include information about discrimination and how health personnel need to keep this in mind.

Health personnel will only use EKG AI if there is a suspicion that the patient may have heart failure. The result from the algorithm is sent back to the EKG archive (ComPACS) at Ahus. The result is displayed to health personnel in a text format in a field next to the patient's EKG reading. Ahus is aware of the risk of notification fatigue in health personnel if too many warnings and too much information appears at the same time. As EKG AI will be used in emergency situations, it is especially important that the information provided is clear and precise. The sandbox project has therefore discussed different ways the prediction can be presented. The prediction can be presented as a percentage, as categories of either "low", "medium" or "high", or a limit could be defined for emergency/non-emergency follow-up. In this process, it will be beneficial to involve health personnel to get insight into what the best way of presenting the results could be. Ahus will explore how to specifically present the prediction in a clinical trial. In a trial, it will be possible to try out the decision support tool in practice on a limited number of patient and with selected health personnel.

---

[22] Hjerte- og karregisteret: Rapport for 2012–2016. From https://www.fhi.no/globalassets/dokumenterfiler/rapporter/2016/hjerte--og-karregisteret.-rapport-for-2012-2016.pdf

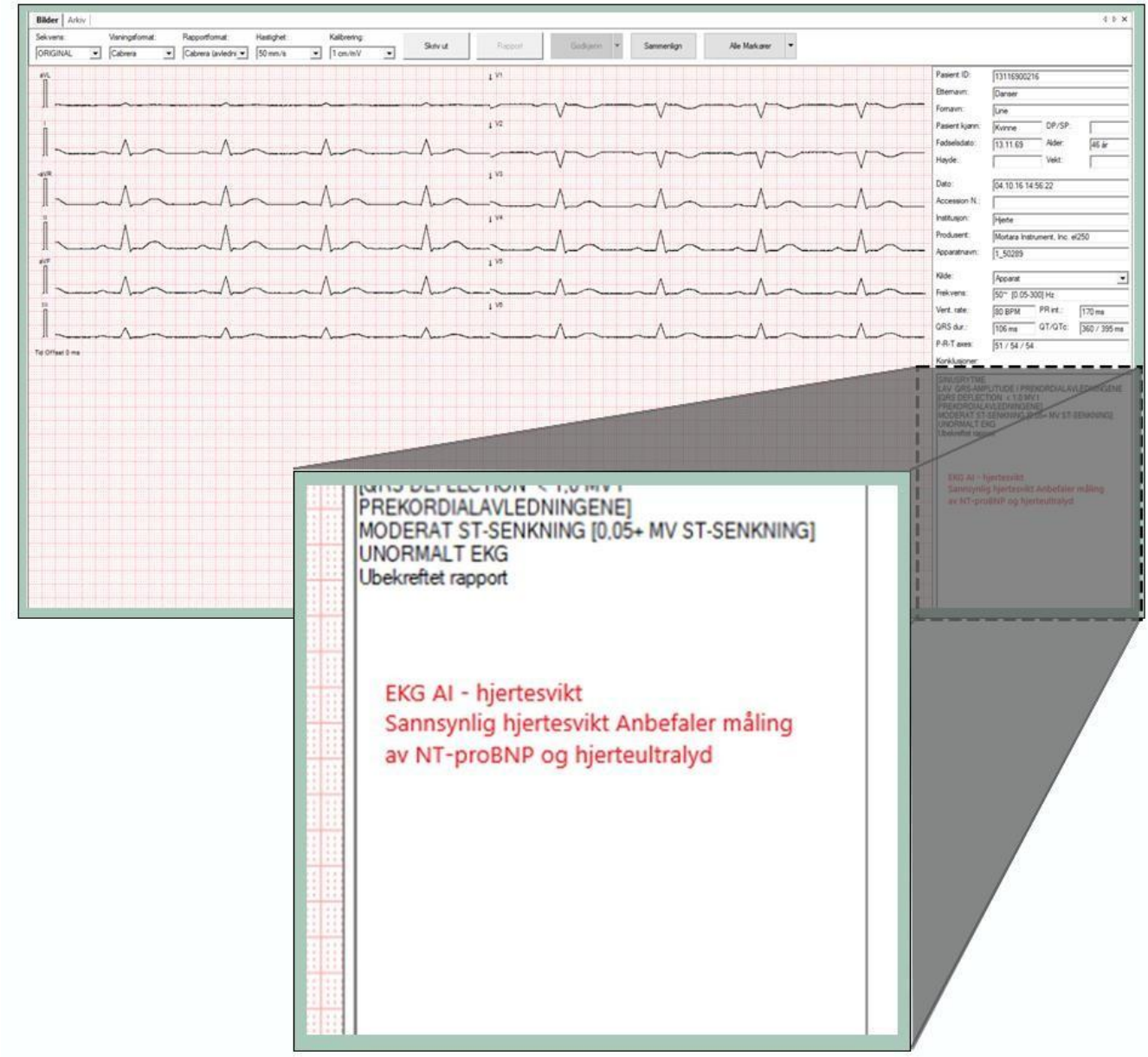


Illustration showing how the results from EKG AI can be presented to health personnel.

As a result of the results from the algorithm being stored in the EKG archive (ComPACS), there is a risk that the result will not be discovered by health personnel immediately. Ahus will therefore consider alternative means of notification, e.g. through monitors at the hospital or by sending a message to the health personnel's work phones. However, this type of notification requires better infrastructure than is currently available at Ahus.

In connection with health personnel training, Ahus will establish procedures and protocols for use of the tool. When the algorithm is developed further, it would be natural for procedures, protocols and the training itself to also be updated. Ahus has an existing non-conformity reporting system, where health personnel report non-conformities that can be used to improve the algorithm in the future.

Technical measure: Monitoring mechanism for continual learning

Over time and with societal changes, EKG AI's predictions will become less accurate. Lower accuracy will occur naturally when a population changes. As an example, new patient groups may emerge, due to the arrival of refugees from a country the algorithm has no data for. When the accuracy is no longer satisfactory, continual learning of the algorithm will be required. Continual learning entails training with new data, testing and validation of the algorithm.

The EKG AI algorithm does not have continuous post-training, which means that its accuracy will not automatically adjust to future changes. Instead, Ahus plans to implement a monitoring mechanism, designed to be triggered when the algorithm's accuracy falls below a pre-defined limit, and the algorithm needs post-training. In order to validate such a limit value, Ahus will conduct a clinical trial while algorithm is being tested in a clinical setting.

In practice, the monitoring mechanism will compare the algorithm's prediction with the diagnosis health personnel set for the patient. This way, it will be possible to assess the degree to which the algorithm's prediction is accurate for the patient's real-life medical condition. The limit value for accuracy will then determine when and if post-training of the algorithm is required.

Going forward

If the project succeeds in developing a good prediction model, the goal is to test it in a clinical setting in early 2024. The next step is to get the algorithm CE-certified and approved by the Norwegian Medicines Control Authority.

Clinical decision-support tools based on artificial intelligence are considered medical-technical equipment and must be approved for use in a clinical setting. An important distinction between ordinary medical-technical equipment and a decision-support tool based on artificial intelligence, is that the latter must be retrained regularly to prevent reductions in accuracy. The Norwegian Medicines Control Authority does not currently have a precedent for approving medical-technical equipment that requires retraining. Ahus has therefore been granted project funding from Live Science Growth House to explore, in consultation with DNV, the possibilities of CE-certifying an algorithm that requires retraining.

This sandbox project has highlighted a potential risk of the EKG AI discriminating against some patient groups that are currently under-represented in the algorithm's data source. Ahus will conduct a clinical trial to determine whether the algorithm's predictions are less accurate for patients with a different ethnic background or has the potential to discriminate on other grounds. The results will indicate whether corrective measures are required.

In the project period, we have discovered that there is no common, baseline method for identifying algorithmic bias. If we had had more time in the project, we would have developed our own method, based on experiences gained in the project period. In addition, it would have been interesting to dive even deeper into the ethical requirements for use of artificial intelligence in the health sector.

The Norwegian Data Protection Authority's regulatory sandbox for responsible artificial intelligence

Office address:
Trelastgata 3, Oslo

Postal address:
PO Box 458 Sentrum
0105 OSLO
Norway

sandkasse@datatilsynet.no
Phone: +47 22 39 69 00

datatilsynet.no/sandkasse
personvernbloggen.no