TIMESPAN

Management of chronic cardiometabolic disease and treatment discontinuity in adult ADHD patients

H2020 – 965381

D7.5. – EDB report (incl. EDAC feedback) on TIMESPAN ethics and data management implementation

Dissemination level	Public
Contractual date of delivery	31 March 2024
Actual date of delivery	05 April 2024
Type	Report
Version	1
Filename	TIMESPAN_Deliverable Report_D7.5
Workpackage	7
Workpackage leader	Patrick Ip (HKU)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965381.

This report reflects only the author's views and the Commission is not responsible for any use that may be made of the information it contains.

Author list

Organisation	Name	Contact information
ORU	Henrik Larsson	Henrik.Larsson@oru.se

Abbreviations

Kommentiert [HL1]: I have updated. there many abbreviations that were not used.

~~**EDAC**~~	~~Ethics and Data Management Advisory Committee~~
ADHD	Attention Deficit Hyperactivity Disorder
AI	Artificial Intelligence
DLNN	Deep Learning Neural Networks
~~**ADHD**~~	~~Attention Deficit Hyperactivity Disorder~~
DMP	Data Management Plan
DPO	Data protection officer
EDAC	Ethics and Data Management Advisory Committee
PI	Principle Investigator
~~**DPO**~~	~~Data protection officer~~
~~**UCN**~~	~~Unique Code Number~~
~~**DLNN**~~	~~Deep Learning Neural Networks~~
~~**DPO**~~	~~Data protection officer~~
~~**EEA**~~	~~European Economic Area~~
~~**SSH**~~	~~Secure Shell~~
~~**ACLs**~~	~~Access Control Lists~~
~~**RSA**~~	~~Rivest-Shamir-Adleman~~
~~**eCRF**~~	~~electronic Case Report Form~~
~~**GPS**~~	~~Global Positioning System~~
~~**EDA**~~	~~Electrodermal Activity~~
~~**AE**~~	~~Adverse Event~~
~~**RPD**~~	~~Raw Physiological Data~~
~~**PPD**~~	~~Processed physiological data~~
~~**MDA**~~	~~Meta data~~
~~**HWD**~~	~~Hardware data~~
~~**CSV**~~	~~Command Separated Values~~

Table of Contents

1. Executive Summary

This deliverable contains the Ethics and Data Management Board report (incl. EDAC feedback) on TIMESPAN ethics and data management implementation. In addition, the report deals with:

- Prevention of misuse of research participant's data (data security measures to prevent unauthorized access and data breach, anonymization techniques, encryption, secure data transfer, controller processor agreement ensuring high security standards (for the remote collection)). Such measures will mitigate the possible risks for the research participants.
- Ethical considerations in relation to AI applications (terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability).
- ~~Next~~ In addition to that the entire report has been evaluated by an independent ethics advisor

Updates D7.5

- The data security section has been updated. More specifically, SUNY (WP6 lead from US) will NOT be provided with access to prescription/electronic health record databases and national registers from Sweden, Denmark, UK, and Hong Kong, as well as from the available cohort studies with data collection from the Netherlands and Estonia. Instead, analyses will be conducted locally under supervision from SUNY.

2. Deliverable report

TIMESPAN's main objective is to advance the management of adult Attention Deficit Hyperactivity Disorder (ADHD) and co-occurring cardiometabolic disease by improving the identification and treatment of individuals with these comorbidities. In this project, we will process sensitive personal data from available (a) prescription/electronic health record databases and national registers; B) available cohort studies with data collection (for details please see MS52 Interim Report to EDAC on newly collected cohort) and newly collected data (Remote measurement technology data via ART-Carma UK and ART-Carma Spain). Extensive work has been and constantly will be devoted to harmonize variables across data sources and to build common protocol(s) and distributed network approach to harmonize study designs, analyses, data cleaning and result outputs across all collaborating sites. This approach is used to adhere to the fact that most data need to stay within each country, cannot be shared in raw form, and must be analysed at the servers of the host university. That is, according to international and national regulations, it is not possible to make data openly accessible. ~~and only some partners are allowed to provide access to the data.~~

TIMESPAN will intensively investigate a huge amount of data. In order to ensure that data is managed properly we ~~will~~ have developed ~~provide~~ an ethical strategy and (FAIR) data management plan (DMP) to allow for maximal transparency, open access/science, usability and reproducibility. Data sustainability will be accomplished by placing all the data management and analyses codes in an online repository (i.e., Github) and by providing a description of how to access raw data of each data source. In addition, we will also maximize transparency and enable future research by presenting aggregated data for all study variables (i.e., made available in the appendix of all publications). We can confirm all data transfers to the UK imply GDPR compliance and compliance of the UK databases.

- For details, please see D7.2. Data Management Plan.
- For details on data storage and management across sites in the newly collected cohort (ART-CARMA) see MS52 Interim Report to EDAC on newly collected cohort – already reviewed by the EDAC

Background:
The purpose of collecting new data from detailed day-to-day monitoring of adult ADHD patients through active and passive are:

- Our first main aim is to obtain real-world data from the patient's daily life on the extent to which ADHD medication treatment and physical activity, individually and jointly, may influence cardiometabolic risks in adults with ADHD. This will provide new insights into disease patterns and help improve the safety and effectiveness of pharmacological (i.e., ADHD medication treatment) and non-pharmacological (i.e., physical activity) interventions for patients with ADHD and co-occurring cardiometabolic disease.
- Our second main aim is to obtain in vivo, real-world data from the patient's daily life on adherence to pharmacological treatment and its predictors and correlates, over a remote monitoring period of 12 months that starts from pre-treatment initiation. The long-term goal is to use these data to improve the management of cardiometabolic disease in adults with ADHD, and to improve ADHD medication treatment adherence and the personalisation of treatment.

Next to that TIMESPAN will also use available data from prescription/electronic health record databases and national registers in Sweden, Denmark, US, Norway, UK, Hong Kong, Iceland and Australia, as well as from the cohort studies with already collected data from Sweden (Lifegene, Swedish Twin registry), the Netherlands (LIFELINE, Trails, Neuroimage), Iceland (SAGA) and Estonia (Estonian Biobank).

1) Prevention of misuse of research participant's data (data security measures to prevent unauthorized access and data breach, anonymization techniques, encryption, secure data transfer, controller processor agreement ensuring high security standards (for the remote collection)). Such measures will mitigate the possible risks for the research participants.

- Access to these data sources have been obtained/are obtained after ethical approval (in the relevant country) and protocol approval (from relevant data source owner).
- Pseudonymized data are then provided to the host (i.e., researcher team at each collaborating site) and data is stored at a secure server at the host university.
- In general, secondary statistical analyses are conducted by the host guided by metadata (for variable harmonization), common protocol(s) and distributed network approach for harmonization of study design, analysis details, data cleaning and result outputs across all collaborating sites
- International and national regulations do not allow making any of the data openly accessible. For example, the Swedish register-data underlying this project contain sensitive personal information and therefore cannot be made openly accessible as they are subject to secrecy in accordance with the Swedish Public Access to Information and Secrecy Act. Researchers may apply for access to the data through the Swedish Research Ethics Boards (www.etikprovningsmyndigheten.se) and from the primary data owners Statistics Sweden (www.scb.se), and the National Board of Health and Welfare (socialstyrelsen.se), in accordance with Swedish law.
- Informed consent:
    - According to national law informed consent is not needed for the available data from prescription/electronic health record databases and national registers in Sweden, US, Denmark-register, Norway, UK, Hong Kong, Iceland, UK and Australia.
    - Informed consent is available for the cohort studies with data collection from Sweden (Lifegene, Swedish Twin registry), the Netherlands (LIFELINE, Trails, Neuroimage), Iceland (SAGA) and Estonia (Estonian Biobank). The Danish iPSYCH cohort data use a system of passive consent together with an easily accessible opt-out option (see further information (only in Danish) at https://nyfoedte.ssi.dk/opbevaring-og-brug-af-proeven).

    - Informed consent is available for the newly collected data in ART-Carma UK and ART-Carma Spain. ART-CARMA has been registered on https://clinicaltrials.gov/.
- Data security:
    - The general approach in TIMESPAN is that all data from prescription/electronic health record databases and national registers and cohort studies with data collection are pseudonymized and all data are stored locally on secure servers at the host university without access to the identity of the individuals. Secure access will require individual investigators to have a user name and password to access the data files. In some WPs, data will be shared with other partners within TIMESPAN. More specifically, Iceland will share data with Sweden. Only coded pseudonymized data will be shared via remote access. The sharing of data within TIMESPAN will comply with the General Data Protection Regulation and any other applicable law or regulation regarding data sharing. The sharing will be provided after data processing agreement and/or European model contract has been approved by all involved. Data Transfer agreement are in place for some of the data sets and for the others data transfer agreement are still under negotiation with the respective legal departments. At each site there will be a study coordinator (Principal Investigator; PI) responsible for data storage and data management. At each site there is a data protection officer (DPO) appointed to safeguard the rights of the research participants (see D7.2. Data Management Plan).
    - Data will be stored locally on a secure server at each host university responsible for a data source. Data will be stored locally 10-30 years at the host university on permanent and secure files following the guidelines for record retention at the host university. Whenever possible, at the end of the project, the data collected within TIMESPAN will be placed in local or national repositories for use by others, according to the ethical procedures of the individual partners.

2) Ethical considerations in relation to AI applications (terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability).

- Ethical considerations in relation to AI applications are considered and further developed in terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability.
- Although prediction models can be extremely useful in clinical settings, they can have the unintended effect of continuing or worsening health care disparities. This problem occurs when a model is developed in a group that is heavily weighted toward one ethnic, economic or another social group. That model might not be valid for other groups that had not been represented in the predictive modelling effort. This issue is especially acute for genomic data because most large samples have been collected from people with European-American ancestry. This problem is exacerbated by the fact that machine learning models are a "black box" that does not allow for easy interpretation of these models make their decisions.

TIMESPAN addresses these issues in several ways.

- Key features of our prediction modelling are leveraging existing high-quality health ~~relevant~~ data from multiple sources, creating novel, AI disease-risk models using deep learning neural networks (DLNNs), and assessing their accuracy, reliability, reproducibility and generalisability across countries, ethnicities and genders.
- For genomic data, we are creating an innovative model that uses adversarial learning to assure that our models learn from valid disease associated genomic features rather than features associated with ethnicity, race or ancestry.
- We will adhere to the FAIR data principles and assure the appropriate use and interpretation of our data along with systematic efforts to reduce health care disparities by clarifying the relevance of our algorithms to both genders and to minority groups in the populations studied.

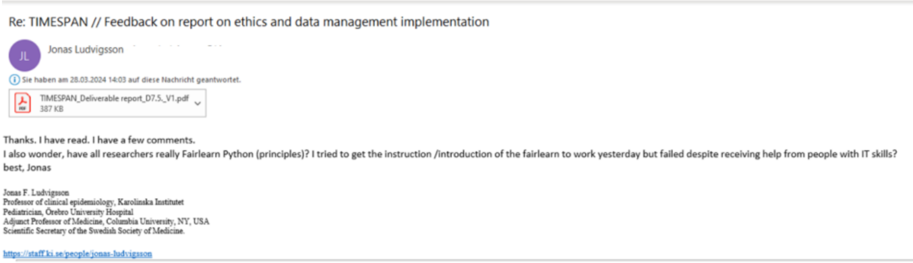
We will adhere to the FAIR data principles using the Fairlearn Python functions for machine learning. Fairlearn is an open-source, community project aimed at improving the fairness of AI systems (https://fairlearn.org/). The Fairlearn functions will be integrated into our workflow to assess fairness metrics for racial, ethnic, gender, immigration status and other disparities. We will also apply Fairlearn disparity mitigation strategies as needed to eliminate any disparities detected in our algorithms.

- Although we cannot completely solve the "black box" problem, we will report feature importance scores, which quantify the effect that each feature has in the decision-making process implemented by algorithms. That will provide some insight into potential biases. For example, if socioeconomic status or sex is an important predictive feature, we will need to do additional modeling of substrata to be sure that such variables are used validly (e.g., as they would be for modeling hypertension) or if they reflect a modeling bias that should be corrected.
- The SUNY site has also led an effort to develop guidelines for reporting machine learning investigations in neuropsychiatry. These standards, which are described in a manuscript submitted for publication, are meant to help researchers avoid errors and misinterpretations, including those that lead to health care disparities.

3. Conclusion

Our Ethics and Data Management Board report (incl. EDAC feedback) describes TIMESPAN ethics and data management implementation as well as a) approaches to prevent misuse of research participant's data and b) ethical considerations in relation to AI applications. The report covers both newly collected data and available data from prescription/electronic health record databases, national registers and cohort studies.

This deliverable report has been reviewed by the TIMESPAN EDAC. All suggested changes have been included.



Re: TIMESPAN // Feedback on report on ethics and data management implementation

Jonas Ludvigsson

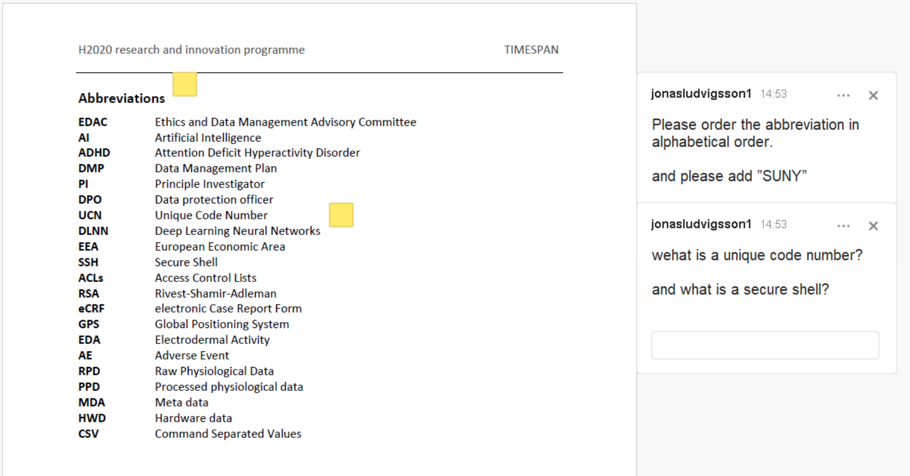
Sie haben am 28.03.2024 14:03 auf diese Nachricht geantwortet.

TIMESPAN_Deliverable report_D7.5._V1.pdf
387 KB

Thanks. I have read. I have a few comments.
I also wonder, have all researchers really Fairlearn Python (principles)? I tried to get the instruction /introduction of the fairlearn to work yesterday but failed despite receiving help from people with IT skills?
best, Jonas

Jonas F. Ludvigsson
Professor of clinical epidemiology, Karolinska Institutet
Pediatrician, Örebro University Hospital
Adjunct Professor of Medicine, Columbia University, NY, USA
Scientific Secretary of the Swedish Society of Medicine.

https://staff.ki.se/people/jonas-ludvigsson



H2020 research and innovation programme TIMESPAN

Abbreviations

EDAC	Ethics and Data Management Advisory Committee
AI	Artificial Intelligence
ADHD	Attention Deficit Hyperactivity Disorder
DMP	Data Management Plan
PI	Principle Investigator
DPO	Data protection officer
UCN	Unique Code Number
DLNN	Deep Learning Neural Networks
EEA	European Economic Area
SSH	Secure Shell
ACLs	Access Control Lists
RSA	Rivest-Shamir-Adleman
eCRF	electronic Case Report Form
GPS	Global Positioning System
EDA	Electrodermal Activity
AE	Adverse Event
RPD	Raw Physiological Data
PPD	Processed physiological data
MDA	Meta data
HWD	Hardware data
CSV	Command Separated Values

jonasludvigsson1 14:53

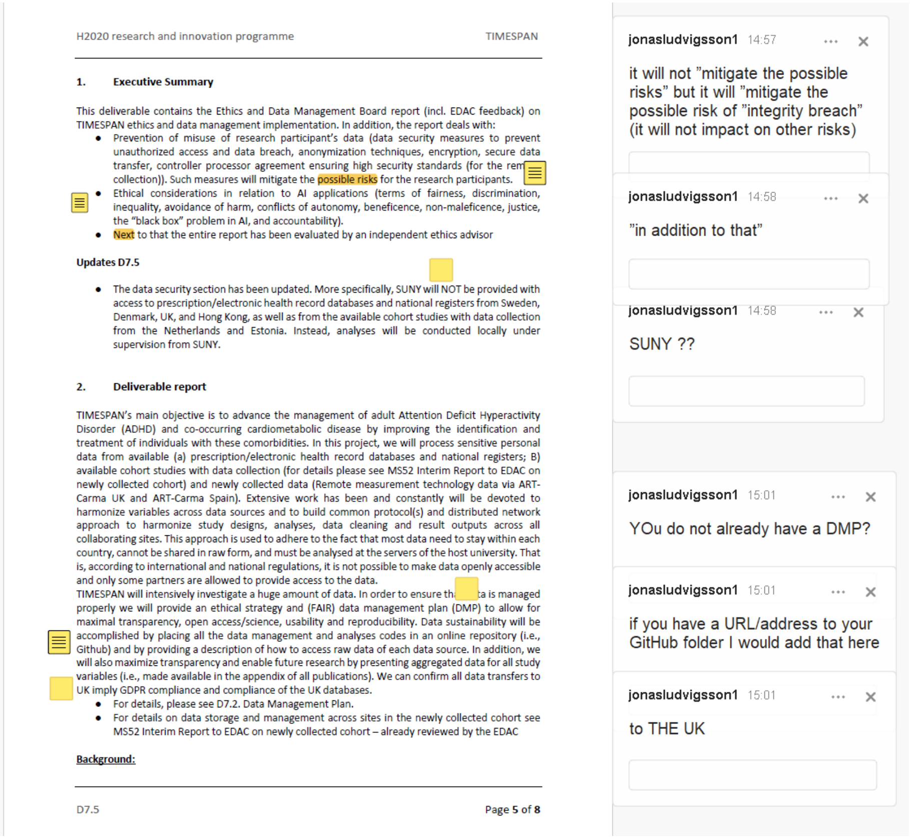
Please order the abbreviation in alphabetical order.

and please add "SUNY"

jonasludvigsson1 14:53

wehat is a unique code number?

and what is a secure shell?



H2020 research and innovation programme TIMESPAN

1. Executive Summary

This deliverable contains the Ethics and Data Management Board report (incl. EDAC feedback) on TIMESPAN ethics and data management implementation. In addition, the report deals with:

- Prevention of misuse of research participant's data (data security measures to prevent unauthorized access and data breach, anonymization techniques, encryption, secure data transfer, controller processor agreement ensuring high security standards (for the rem collection)). Such measures will mitigate the possible risks for the research participants.
- Ethical considerations in relation to AI applications (terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability).
- Next to that the entire report has been evaluated by an independent ethics advisor

Updates D7.5

- The data security section has been updated. More specifically, SUNY will NOT be provided with access to prescription/electronic health record databases and national registers from Sweden, Denmark, UK, and Hong Kong, as well as from the available cohort studies with data collection from the Netherlands and Estonia. Instead, analyses will be conducted locally under supervision from SUNY.

2. Deliverable report

TIMESPAN's main objective is to advance the management of adult Attention Deficit Hyperactivity Disorder (ADHD) and co-occurring cardiometabolic disease by improving the identification and treatment of individuals with these comorbidities. In this project, we will process sensitive personal data from available (a) prescription/electronic health record databases and national registers; B) available cohort studies with data collection (for details please see MS52 Interim Report to EDAC on newly collected cohort) and newly collected data (Remote measurement technology data via ART-Carma UK and ART-Carma Spain). Extensive work has been and constantly will be devoted to harmonize variables across data sources and to build common protocol(s) and distributed network approach to harmonize study designs, analyses, data cleaning and result outputs across all collaborating sites. This approach is used to adhere to the fact that most data need to stay within each country, cannot be shared in raw form, and must be analysed at the servers of the host university. That is, according to international and national regulations, it is not possible to make data openly accessible and only some partners are allowed to provide access to the data.

TIMESPAN will intensively investigate a huge amount of data. In order to ensure th ta is managed properly we will provide an ethical strategy and (FAIR) data management plan (DMP) to allow for maximal transparency, open access/science, usability and reproducibility. Data sustainability will be accomplished by placing all the data management and analyses codes in an online repository (i.e., Github) and by providing a description of how to access raw data of each data source. In addition, we will also maximize transparency and enable future research by presenting aggregated data for all study variables (i.e., made available in the appendix of all publications). We can confirm all data transfers to UK imply GDPR compliance and compliance of the UK databases.

- For details, please see D7.2. Data Management Plan.
- For details on data storage and management across sites in the newly collected cohort see MS52 Interim Report to EDAC on newly collected cohort – already reviewed by the EDAC

Background:

D7.5 Page 5 of 8

jonasludvigsson1 14:57

it will not "mitigate the possible risks" but it will "mitigate the possible risk of "integrity breach" (it will not impact on other risks)

jonasludvigsson1 14:58

"in addition to that"

jonasludvigsson1 14:58

SUNY ??

jonasludvigsson1 15:01

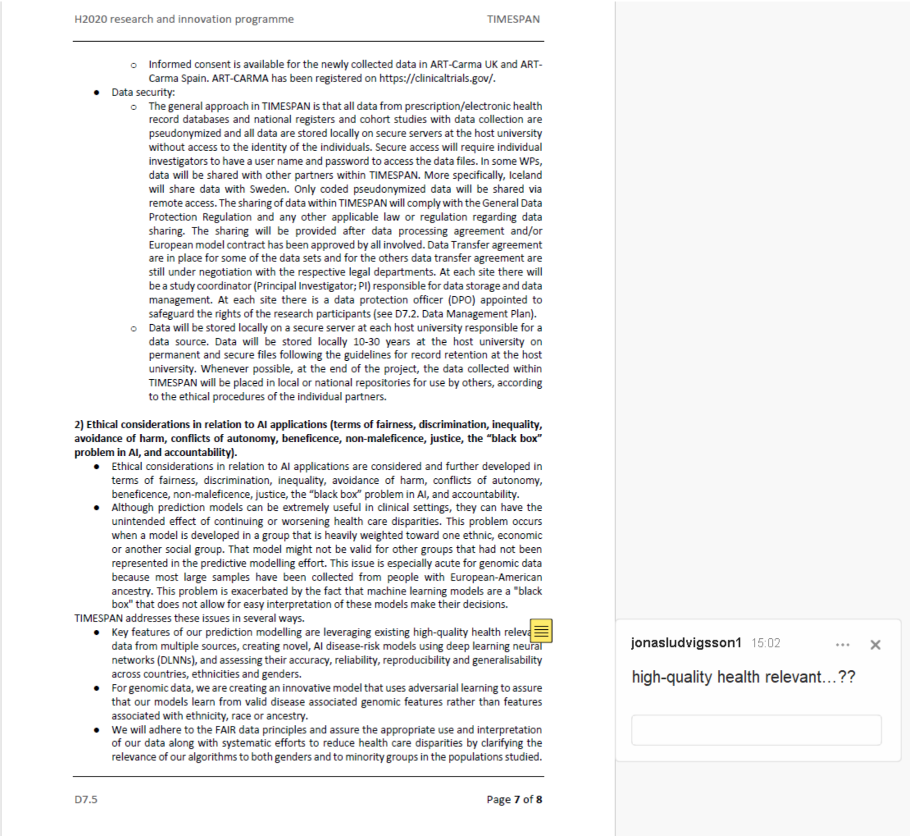
YOu do not already have a DMP?

jonasludvigsson1 15:01

if you have a URL/address to your GitHub folder I would add that here

jonasludvigsson1 15:01

to THE UK



- Informed consent is available for the newly collected data in ART-Carma UK and ART-Carma Spain. ART-CARMA has been registered on https://clinicaltrials.gov/.
- Data security:
    - The general approach in TIMESPAN is that all data from prescription/electronic health record databases and national registers and cohort studies with data collection are pseudonymized and all data are stored locally on secure servers at the host university without access to the identity of the individuals. Secure access will require individual investigators to have a user name and password to access the data files. In some WPs, data will be shared with other partners within TIMESPAN. More specifically, Iceland will share data with Sweden. Only coded pseudonymized data will be shared via remote access. The sharing of data within TIMESPAN will comply with the General Data Protection Regulation and any other applicable law or regulation regarding data sharing. The sharing will be provided after data processing agreement and/or European model contract has been approved by all involved. Data Transfer agreement are in place for some of the data sets and for the others data transfer agreement are still under negotiation with the respective legal departments. At each site there will be a study coordinator (Principal Investigator; PI) responsible for data storage and data management. At each site there is a data protection officer (DPO) appointed to safeguard the rights of the research participants (see D7.2. Data Management Plan).
    - Data will be stored locally on a secure server at each host university responsible for a data source. Data will be stored locally 10-30 years at the host university on permanent and secure files following the guidelines for record retention at the host university. Whenever possible, at the end of the project, the data collected within TIMESPAN will be placed in local or national repositories for use by others, according to the ethical procedures of the individual partners.

2) Ethical considerations in relation to AI applications (terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability).

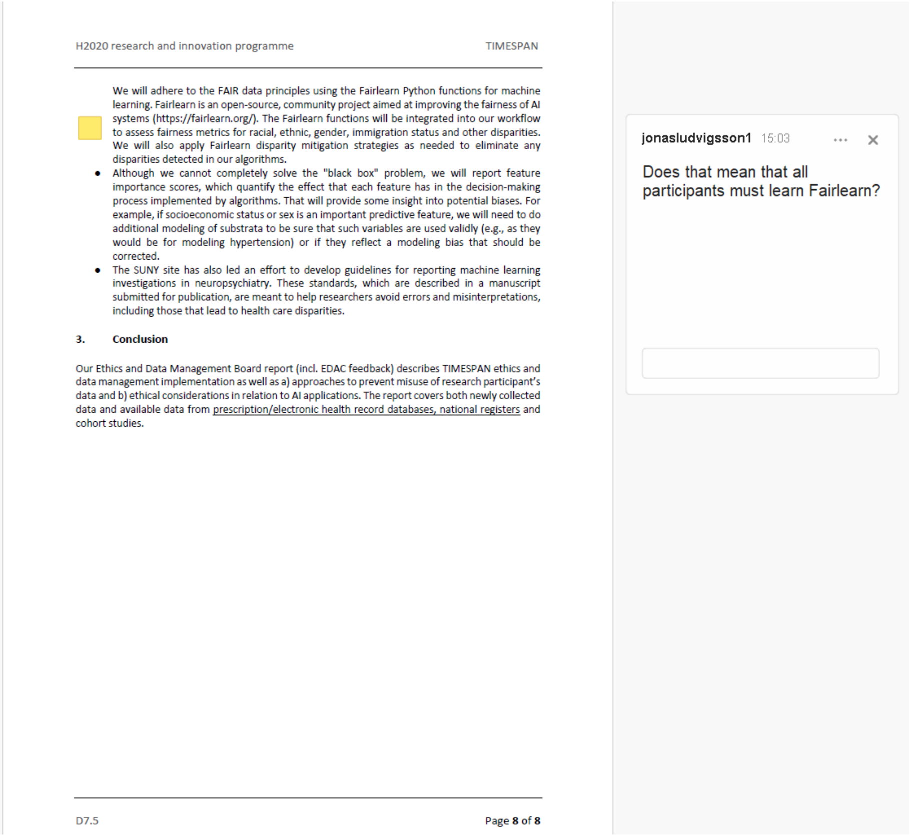
- Ethical considerations in relation to AI applications are considered and further developed in terms of fairness, discrimination, inequality, avoidance of harm, conflicts of autonomy, beneficence, non-maleficence, justice, the "black box" problem in AI, and accountability.
- Although prediction models can be extremely useful in clinical settings, they can have the unintended effect of continuing or worsening health care disparities. This problem occurs when a model is developed in a group that is heavily weighted toward one ethnic, economic or another social group. That model might not be valid for other groups that had not been represented in the predictive modelling effort. This issue is especially acute for genomic data because most large samples have been collected from people with European-American ancestry. This problem is exacerbated by the fact that machine learning models are a "black box" that does not allow for easy interpretation of these models make their decisions.

TIMESPAN addresses these issues in several ways.

- Key features of our prediction modelling are leveraging existing high-quality health releva data from multiple sources, creating novel, AI disease-risk models using deep learning neural networks (DLNNs), and assessing their accuracy, reliability, reproducibility and generalisability across countries, ethnicities and genders.
- For genomic data, we are creating an innovative model that uses adversarial learning to assure that our models learn from valid disease associated genomic features rather than features associated with ethnicity, race or ancestry.
- We will adhere to the FAIR data principles and assure the appropriate use and interpretation of our data along with systematic efforts to reduce health care disparities by clarifying the relevance of our algorithms to both genders and to minority groups in the populations studied.

jonasludvigsson1 15:02

high-quality health relevant…??



H2020 research and innovation programme TIMESPAN

We will adhere to the FAIR data principles using the Fairlearn Python functions for machine learning. Fairlearn is an open-source, community project aimed at improving the fairness of AI systems (https://fairlearn.org/). The Fairlearn functions will be integrated into our workflow to assess fairness metrics for racial, ethnic, gender, immigration status and other disparities. We will also apply Fairlearn disparity mitigation strategies as needed to eliminate any disparities detected in our algorithms.

- Although we cannot completely solve the "black box" problem, we will report feature importance scores, which quantify the effect that each feature has in the decision-making process implemented by algorithms. That will provide some insight into potential biases. For example, if socioeconomic status or sex is an important predictive feature, we will need to do additional modeling of substrata to be sure that such variables are used validly (e.g., as they would be for modeling hypertension) or if they reflect a modeling bias that should be corrected.
- The SUNY site has also led an effort to develop guidelines for reporting machine learning investigations in neuropsychiatry. These standards, which are described in a manuscript submitted for publication, are meant to help researchers avoid errors and misinterpretations, including those that lead to health care disparities.

3. Conclusion

Our Ethics and Data Management Board report (incl. EDAC feedback) describes TIMESPAN ethics and data management implementation as well as a) approaches to prevent misuse of research participant's data and b) ethical considerations in relation to AI applications. The report covers both newly collected data and available data from prescription/electronic health record databases, national registers and cohort studies.

D7.5 Page **8** of **8**

jonasludvigsson1 15:03

Does that mean that all participants must learn Fairlearn?