|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ITU Logo | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2017-2020 | | FG-AI4H-H-005 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Brasilia, 22-24 January 2020 | |
| **DOCUMENT** | | | | |
| **Source:** | | TSB | | |
| **Title:** | | Summary – DASH/DAISAM Workshop (Berlin, 8-9 January 2020) | | |
| **Purpose:** | | Discussion | | |
| **Contact:** | | Luis Oala Fraunhofer HHI Germany | | Email: luis.oala@online.de |
|  | | Marc Lecoultre BI GPS Switzerland | | Email: marc.lecoultre@bigps.ch |
|  | | Pradeep Balachandran eHealth Consultant India | | Email: pbn.tvm@gmail.com |
|  | | Monique Kuglitsch Fraunhofer HHI Germany | | Email: [m](mailto:pbn.tvm@gmail.com)onique.kuglitsch@hhi.fraunhofer.de |
|  | | Eva Weicken Fraunhofer HHI Germany | | Email: [e](mailto:pbn.tvm@gmail.com)va.weicken@hhi.fraunhofer.de |
|  | | Lars Roemheld health innovation hub Germany | | Email: lars.roemheld@hih-2025.de |
|  | | Johannes Starlinger 37 binary Germany | | Email: johannes.starlinger@37binary.com |
|  | | Ferath Kherif CHUV Switzerland | | Email: Ferath.Kherif@chuv.ch |
|  | | Workshop participants as per [this list](https://docs.google.com/spreadsheets/d/1u3p5QrqkArL8_tJ8I1O5_j3qYIeycYLP0TD0siWmfM4/edit?usp=sharing) | | Email: - |

|  |  |
| --- | --- |
| **Abstract:** | This document provides an overview of the contents and results of the 1st DASH/DAISAM Workshop that took place in Berlin between January 8 and 9, 2020. A condensed and edited report is planned for later publication. |



Table of Contents

[1. Program 2](#_Toc30172323)

[1.1 Day 1 (January 8) 3](#_Toc30172324)

[3](#_Toc30172325)

[1.2 Day 2 (January 9) 4](#_Toc30172326)

[2 Notes from the Work Sessions 5](#_Toc30172327)

[2.1 Tests WS 5](#_Toc30172328)

[1.1 Data WS 51](#_Toc30172329)

[1.2 Assessment Platform WS 83](#_Toc30172330)

[3 Result Presentations from the Work Sessions Chairs 103](#_Toc30172331)

[3.1 Tests WS 103](#_Toc30172332)

[104](#_Toc30172333)

[3.2 Data WS 105](#_Toc30172334)

[3.3 Assessment Platform WS 106](#_Toc30172335)

[4 Workshop Participant Volunteers for Contribution to FG-AI4H Deliverables 108](#_Toc30172336)

# Program

Original file: <https://docs.google.com/spreadsheets/d/1keTvJXLwlKF2HUVA9ogrA1uDDiqO_9TiMtw5XkCIU54/edit?usp=sharing>

## Day 1 (January 8)

## 

## Day 2 (January 9)

# Notes from the Work Sessions

## Tests WS

Original file: [*https://docs.google.com/document/d/1-R7PwlQz9lQA\_xEBpRTNNiVkEEVb\_kO-InnghAYd6tE/edit?usp=sharing*](https://docs.google.com/document/d/1-R7PwlQz9lQA_xEBpRTNNiVkEEVb_kO-InnghAYd6tE/edit?usp=sharing)

Explainable Convolutional Neural Networks for Analyzing Structural MRI Data in Neurological Diseases

*Prof. Dr. rer. nat. Kerstin Ritter, (Germany), H*

Junior professor Computational Neuroscience - Charité Berlin

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| [YOUR NOTES HERE] |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| How can we automate current screening or diagnostic practices which solely depend on the expertise of medical specialists? | Can AI play a supportive role - Augmentation or Replacement roles or imitate the diagnostic process of radiologists? | [YOUR NOTES HERE] |
| Why most of the current ML applications in neuroimaging are seen concentrated to ‘Diagnostics’ practice in comparison to other processes such as treatment response,prognosis, subtype identification, symptom identification, etc?  Because of lack of data available for processes such as treatment response,prognosis, subtype identification, symptom identification | Provision of more open data bases required in the case of the mentioned tasks- treatment response,prognosis, subtype identification, symptom identification, etc  Some of the available Neuroimaging data bases include- DELCODE, PPMI, HCP,IMAGN,ADNI,ABCD,UK Biobank,etc |  |
| Available Neuroimaging database sizes are very small and are in the order of few hundreds and thousands only. How can we resolve this issue? | Data in-sufficiency problem can be addressed to a large extent by schemes such as - Data Augmentation , Transfer Learning, etc. |  |
| In case of specific use cases belonging to PTSD, Pain, Anxiety, Depression, ADHD, Autism, etc., it is seen that only very less data is available compared to that of Alzheimer's disease | Provision of more open data bases required in the case of mentioned use cases |  |
| Why Deep Learning Preferred in comparison to Standard Machine Learning in neuroimaging? | In Deep Learning we have the advantage of not taking into account any prior ‘ feature extraction’ process  Deep Learning Networks are optimized for data coming from multiple arrays( images and videos) and incorporates different types of layers such as convolution, poling and non linearity and closely mimics the functioning of human primary visual cortex  Recent study reports reveal that state-of-the-art CNNs are capable of matching and in some cases outperforming human performance levels in tasks related to medical imaging such as classification, segmentation, etc. |  |
| How to address the issue of ‘ Biased Medical Databases’? How can we detect Bias? | Provide Good documentation (e.g Model Cards)  Extend the evaluation of Accuracy in case of subgroups like age,gender, ethnicity, etc. | Differentiated analysis based on Bias relevant subgroups may be performed |
| Deep Learning as a Black Box phenomena- “Interpretability Vs Performance” tradeoff- How can we improve explainability of Deep Learning Models? | One method is by generation of ‘ heat-maps’ based on input data for network decision and model validation.  Train CNN Model and apply to Test Data. Then produce heatmap for each subject in the Test Data Set  Augment the AI Model result ( e.g probability of a disease condition ) with heatmap of test data which will aid explainability through visual comparison of heatmaps and thus helps build trust in the AI based procedure |  |
| How can we address the complex question of predicting disease activity in Multiple Sclerosis patients based on baseline data? | Collaborative data sharing required for provision of different databases from different hospitals |  |
| How much feasible is to translate AI solutions into clinical routines? How can we mitigate the existing misconceptions about perceived ML & Hospital constraints | Need to provide proper awareness to mitigate the existing misconceptions that include  ML needs lot of data  ML prefers standardized data  ML depends on Gold Standard of labels  ML is computationally expensive  ML requires methodological expertise  ML is perceived to be a black box  Need to have the right kind of policies and mechanisms to solve and set right the existing issues in hospitals that includes aspects related to:  Data protection  Heterogeneous & unstructured data  Changing diagnostic guidelines  Unhealthy competition with technology companies and ML experts  Inadequate infrastructure  Lack of transparency and responsibility |  |
| ‘ Prevalence in test database’ may not reflect ‘ real-world prevalence’ . How to address this scenario? | By provisioning of Combined databases, Federated learning(e.g. Pooling data from primary, secondary and tertiary hospitals) | Subgroup analysis- but which subgroups are defined? Need to take comorbidities into account |
| Defining healthy controls- e.g. general population or patients with symptoms that do not turn out to have condition | Need for Standard definition for‘ ‘ Healthy Controls’ |  |
| Is it possible to come up with AI Models that can be fairer than the diagnostic guidelines?  Complexity of diagnostic pathways, which may rely on guidelines but may also rely on intuition. Tests e.g. biochemical are biomarkers, that the physician then uses to make diagnosis- these are separate tasks. AI model could create probability which is then used to make a diagnosis.  Also physicians have acumen of dealing with missing data( based on relevance), where as ML models cannot compromise too much on missing data. | Drawing analogy between ‘ Lab Measurement based Diagnosis ‘ and ‘ Diagnostic AI System’  Bio markers measure a certain value and produce a measurement result.and the actual diagnosis on this biomarker is done by the physician. Legally the diagnosis has to be done by a physician.  'Measurement result generation ' from a lab test and 'diagnosis' have to be seen as 2 seperate taks. e.g the classification task done by an AI system, where it produces a probability value, can be seen similar to the biomarker result from a lab test. But the diagnosis which is based on this probability value be actually done by the physician. |  |
| Data available for training algorithms may be incomplete |  |  |
|  |  | There are scenarios where simple and cost effective methods ( e.g. statistical analysis based ) can be recruited as alternatives to so called EXPENSIVE ‘Gold standards’ . Also some of the current ‘Gold Standards ‘ may not be strictly called so because they have their inherent failure rates. |
|  |  | In the case of two different AI systems having the same architecture but set with different initial values, how consistent would the results be ? |
|  |  | What appropriate metric to be optimized for while designing the algorithm that is primarily based on screening? Typically for higher ‘sensitivity’ and desired level of ‘specificity ‘. But then the specifications might be different when it is optimized for a Gold Standard |
|  |  |  |

ML for Mammography & Inspecting Adversarial Examples Using the Fisher Information

*Dr. rer. nat. Jörg Martin, (Germany), H*

Researcher - Physikalisch Technische Bundesanstalt (PTB)

Information about the topic: https://doi.org/10.1016/j.neucom.2019.11.052

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| [YOUR NOTES HERE] |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| [YOUR NOTES HERE] | [YOUR NOTES HERE] | [YOUR NOTES HERE] |
| How to arrive at optimal dose for optimal image quality in AI for Mammography applications? |  |  |
| ..How do you estimate image quality? | Four-Alternative Forced Choice (AFC) Test is the state of the art | ... |
| What are some of the drawbacks of current automatic image quality estimation processes?  Cumbersome preprocessing due to-Phantom Alignment, Individual Cell Analysis, Disc localization ,etc. | By recruiting End-to-End Machine Learning method( e.g CNN)  Advantages of this CNN method are:  Regression net with multiple layers  Contrast-Detail estimation with single image  Random subsampling  ‘Incremental Learning’ based on extended training data and use of transfer learning  Good quality prediction with less uncertainty  Merits of ML based approach compared to standard method :  More Flexible  More Handy  More Robust  Less cumbersome  Major Threat: Adversarial sample attack | ... |
| Which are the typical adversarial attacks( input perturbation) techniques ?  White box attack ( need access to network architecture information) and Black box attack ( more sophisticated)  Zero box attacks( can cause small perturbations without system access) |  |  |
| Why Testing against adversarial inputs are important? | It acts as a ‘Test for Sensitivity’ against system modification  Resilience to adversarial attacks is a good performance parameter |  |
| Which are the adversarial attack defence mechanisms used? | Detect adversarial examples  Include adversarial examples in training data set |  |
| How to detect adversarial inputs  If input is high dimensional, in most cases the adversarial input is imperceptible.  The criticality of the problem is that  It throws the wrong output with very high confidence  Issue becomes transferable-if you design an adversarial example to fool an AI system, it also works for similar AI systems | ‘ Fisher information ’ ( *a statistical quantity that measures informativity of data*) based decision logic can be used to detect adversarial inputs  Large values of ‘ Fisher information’ indicate the presence of adversarial inputs  ‘ Fisher information sensitivity’ helps to visualize the reason for detection of adversarial input i.e it helps to visualize which regions of the input are responsible for high value of Fisher information |  |
| How feasible is the computation of Fisher information ? | Through appropriate transformation methods, it is possible to make Fisher information computation feasible and make it scale linearly with the network size |  |
| How to deal with Cyber attacks? | Risk mitigation can be carried out by  Detection of inputs from unknown sources  Detection of brittle features of data  Detection of system data access attempts/ request by external programs |  |
|  |  | ‘ Abstract Interpretation’ (software design method) as another emerging and robust defence strategy - which looks at defining strong boundary relations between specific data cloud and its corresponding output |
|  |  | How to deal with the problem of -Crafting input patterns by manipulation of decision boundaries |
|  |  | Neural Networks are good at exploiting the patterns in the data. So If there is some brittle non-robust features ( small change in this feature creates a big spike in the output) and hence this has to be regarded as more of a ‘ data problem’ rather than one associated directly to the model. This calls for careful scrutiny of data to have reasonable expectations on the model. |
|  |  | Creation of adversarial synthetic data set based on the model and on the data available to close the gaps around the decision boundaries |
| This session topic relates to the deliverable- “ AI Technical Test Metric Specification “ |  |  |

Applicability of AI in Aiding Pathology Diagnosis in Low and Middle Income Countries

*Derrick Bary Abila, (United Kingdom), H*

Postgraduate Student - University of Manchester

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| [YOUR NOTES HERE] |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| [YOUR NOTES HERE] | [YOUR NOTES HERE] | [YOUR NOTES HERE] |
| Pathology (both Hysto and Cyto cases)sample processing consists of 3 phases namely- Pre-analytical, Analytical and Post-analytical phases  Pre-analytical: includes factors related to acquiring, handling transporting patient specimen before actual analysis  Analytical : includes factors related to Test Platform and Testing Process  Post-analytical: Includes interpretation of test results by medical expert to guide patient management |  |  |
| Access to pathology services-Challenge in LMICs  Typical Ratio of ( Anatomic Pathologists to Patient Population) stands as low as 1: 1000,000 as compared to 1: 50 as seen in High Income Countries  Other allied challenges include  Inadequate infrastructure, Inadequate Personnel  Inadequate professional Training opportunities  Insufficient funding for procurement of basic lab materials and accessories | Can AI based systems bridge this gap-in the appropriate role of Decision Support Systems in pathology diagnostics  Enablers: How best to reach to Government Ministries of Health ( especially in African Countries) and convince them for adoption of new technologies and advocating the need for a good governance structure for implementation of these projects | ... |
| Considerations for AI solution to pathology | By strict AI technology assessment needed as a precursor to implementation . Need to evaluate potential risk /harm against potential benefits  By adding more ‘explainability’ features to AI solutions  By assessing the effects of outcomes of AI solutions to patients. More clinical studied needed to assess the impact of these AI technologies  By ways to reduce Bias of AI systems  By ways to see how can models be made transferable from one setting to another  Coming with effectiveness criteria, standard Operating Procedures(SOP) | ... |
| Need to come up with criteria for ‘ measuring the ‘effectiveness’ of AI technologies. Shall include ‘regulatory aspects also | The ‘ effectiveness measurement criteria’ may include the following:  Should provide results for a specified clinical problem to guide clinical decisions, for monitoring disease status or response to therapy or to collect data for disease surveillance  Test results must be available in a time frame that will guide clinical decision making  Tests must be easy to perform and results easy to interpret and communicate  Manufacturer’s claims regarding test performance characteristics must be independently verified.  Test platform must be usable and stable in locations of intended use  Test platforms must meet procurement requirements for supply chain,maintenance, availability of quality control standards,durability and stability in variable environmental conditions  Test costs must be affordable in locations of intended use |  |
| AI solution compliance to different operational settings | Need for Generalized and Adaptable AI Model development |  |
| How to handle voluminous historical data in non electronic formats? | Need for digitization of historical data  Provision for data storage and allied information systems  Enablers: The East African Public Health Laboratory Network has spearheaded a program to facilitate the sharing of laboratory data among different sites |  |
| How to maintaining Quality Management systems in Laboratories?  Issues:  Majority of the labs are not accredited and do not follow minimal standards of laboratory practice  Increase in magnitude of false positives and false negative results | By setting forth proper Accreditation Systems that conformance to international standards and also Quality Management Systems to ensure that quality data is used for AI model development as well as the results obtained from model are reliable and consistent  Enablers: Initiative called African Laboratory Medicine Network drives a quality management adoption program in African countries  Also National Drug Authority in Uganda oversees the regulation of new technologies in the health sector |  |
| How to improve adoption of AI Systems among healthcare decision makers? | By providing adequate awareness and education for stakeholders | Since different medical subspecialties vary in terms of expertise and specific workflow languages used, the AI interpretational skills needed by different subspecialties actors also vary accordingly |
|  |  | How to assess the feasibility of different AI deployment settings or models representative of different LMIC scenarios |
|  |  | Improvement in infrastructure is a necessary requirement to participate in successful AI adoption and deployment. But optimization of infrastructure can be brought in by designing suitable service delivery mechanisms like the Hub-Spoke Model to connect large number of distributed member laboratories(Spokes) with regional referral labs ( who will act as Hubs) |
| This session topic relates to the deliverables- “ AI4H-Scaleup and Adoption” & “ Data Handling” |  |  |

Can Assessment of AI-based Medical Devices Learn From Standardization and Round Robin Tests for Clinical Chemistry?

*Prof. Dr. rer. nat. habil. Frank Klawonn, (Germany), H*

Professor/Group Leader - Ostfalia University of Applied Sciences/Helmholtz Centre for Infection Research

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Round robin trials- calibration tests- send copies of the same samples to different labs to compare results  Sometimes artificial samples  Sometimes no gold standard/ ground truth- e.g. result may not be quantifiable  Youden plots- variance and bias  Point of care testing  Fast results- less strict requirements than lab devices |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| [YOUR NOTES HERE] | [YOUR NOTES HERE] | [YOUR NOTES HERE] |
| How the lab laboratory values are used ? | Important concept of reference intervals ( "normal range" & conceptually equivalent to Decision Boundaries) for a lab parameter is defined as the range of values between 2.5 % and 97.5 % quantile of a healthy population  Since reference intervals are derived from ‘healthy population’ data, it is important to have a standard definition for “ healthy population” | ... |
| ’ Reference Intervals’ estimation in clinical lab diagnostics varies due to bio-marker perturbations. Indirect measurements- variation due to differences of measurement devices, chemicals used ,sample processing technique,etc and hence it complicates the evaluation of such devices . Is there a remedy?  Reference Intervals can depend on age, sex,measurement procedure, etc. So how to get a good estimate of extreme quantiles, | Round Robin Trial is normally employed ( Sends copies of same samples to different laboratories to compare the results and to assess the quality of laboratories. Can be considered as calibration tests  Test criteria is usually a maximum difference between repeated measures and maximum deviation from a Gold Standard or Average Value  Typical approaches:Youden Plots and Passing-Bablok Regression  Larger sample sources are required  ISO 15189 recommends that each Lab should compute and maintain their own ‘ Reference Intervals’ | ... |
| What are the some of the constraints of Round Robin Trail based Laboratory quality assessment method?  For some lab parameters, it is not possible to send real ((human) sample and therefore artificial samples are also mixed  For some lab parameters, there is no ‘Gold Standard’ specification |  |  |
| How are the lab values measured? | Guided by Laboratory Diagnostic Pathways |  |
| Laboratory Test Vs POCT devices -Merits and Demerits | Laboratory Test Devices (expensive) -Gives high precision results but processing and waiting time delays are large. I.e how to get precise information in reasonable time  Point-of-Care-Test Devices (POCT)( less expensive) - gives immediate results but with less precision. More applicable to ‘emergency’ situations i.e how to get ‘fast and gross’ information  Quality assessment of POCT devices( less reliable) is less strict compared to Laboratory Test Devices( more sophisticated)  Measurements from Lab Tests and POCTs must be distinguishable |  |
| How to address inter-lab variability of different lab values- How to correctly assess the variations and make it more comparable? | Standardisation / Normalization schemes Data harmonization and interoperability iseven applicable to units under the same hospital or healthcare institution |  |
| Identification and Specification of dependent variables for different AI methods | AI based methods should clearly specify the ‘target population’ for lab measurements because of dependencies with respect to age, sex, etc. |  |
|  |  | To what extent are Round Robin techniques extensible / applicable in case of multiple AI models for the same task solution in comparison to Laboratory processes? |
|  |  | Existing Infrastructure and Standardisation procedure of Round Robin techniques for testing can be leveraged to AI application domain. Now what we need to look for is the procedure specific to AI based method |
|  |  | Since AI based approach also has the objective of how to assess the consistency of accuracies, this can be closely associated with a ‘quality assurance framework ‘ that deals with the similar settings- have inputs, measurement procedure, and outputs produced. So the existing framework can be appropriately extended ( leveraging federated infrastructure and standardization procedures) from laboratory test domain to AI based test domain |
|  |  | How can we leverage the concept of ‘artificial data set’ creation in laboratory test domain to augment the need for or say potential methods to generate limited Test data, more Test data, new Test data , etc.in AI domain? |
| In Round Robin Trial, the core interest is to see that the measurement values agree irrespective of the multitude of settings and irrespective of what conclusions are made out of it.  In comparison, in AI case, we need to investigate further on at what all levels ( at measurement accuracy levels or at conclusion agreement levels, etc.) we need to achieve consistency. |  | Like Laboratory Tests , we need to set up a Standardized Assessment Procedure for AI based Tests |
|  |  | 2 salient aspects of Round Robin Test Vs AI based Tests-  Periodic testing of diagnostic procedure  Comparability- are you able to compare the results of different AI models and draw meaningful conclusions- probably in terms of a match between 2 different AI having the same set of input and output specifications  If quality control mechanism is such that the AI system is ‘Locked and Deterministic’, there may be no need to have a periodic assessment of that AI system as it assures reproducibility of results |
| This session topic relates to the deliverable- “Training and Test Data Specification “ |  |  |

Novel Decision-Informed Metrics for Medical AI Validation

*Federico Cabitza PhD, (Italy), R*

Associate Professor - Università degli Studi di Milano-Bicocca

Information about the topic: https://arxiv.org/pdf/1910.09246.pdf

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Technical details and lots of diagrams in our arxiv paper (submitted to impacted literature): https://arxiv.org/pdf/1910.09246.pdf |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Many metrics and measures to assess the performance of predictive (classification) models embedded in Medical AI (MAI) exist and they have been proposed before the AI community and the medical community to understand the value of prospective ML-based decision support tools.  Besides their number, these metrics are either trivial (standard accuracy), or prone to bias/distortion (e.g., class unbalance), or to misunderstanding (e.g., AUROC, log loss).  Generally speaking, medical doctors don’t understand these measures (besides accuracy). Moreover, these measures are not practice-aware (or informed by medical practice).  In short, we need a simple (one number), intuitive notion of “how good a decision support is”. | To address the above problem, we devised a novel formulation for the expression of accuracy: H-accuracy (Ha).  Essentially, Ha is a balanced, chance-adjusted, weighted (both class- and case-wise) form of accuracy that, for some parametrization, is **equivalent** to either balanced accuracy or standardized Net Benefit and satisfies a number of properties (see article for more details).  Its formulation has been designed in collaboration with expert diagnosticians from three medical fields: the main element of novelty lies in the fact that it considers \***case complexity**\* (aka discriminative difficulty) and weights accuracy accordingly. | You may want to optimize ML models towards either simple cases (e.g., screening, primary healthcare) or complex cases (e.g., secondary healthcare settings, second opinion scenarios, …): H-accuracy allows to “weight” the assessment of model performance according to either preferences by means of the d(x) function and “p” parameter.  Pi-representativeness does not assess biases, but it can be used to this aim, if it used to understand whether the training/test sets are “similar” enough to a reference population dataset (vs gender bias, racial bias, sampling bias…). |
| 3 novel metrics are proposed to evaluate the 'pragmatic validity' of a model w.r.t a Benchmark Test Dataset for a specific discriminative task and these metrics can guarantee merits beyond the typical ' statistical validity' norm used for for model validity evaluation.  It looks as at a data driven approach to assess 'human perceived complexity' and taking it into account for evaluating AI model validity.  In short, an 'accuracy' metric( H-accuracy) , a 'representativeness' metric(Pi-Representativeness) and a 'Robustness' metric(Ratio of H-accuracy and Pi-Representativeness) are proposed. | H-accuracy (Ha) : a novel formulation of accuracy to represent the practical value and to assess the machine classification model with respect to any Benchmark Test Dataset for which we collect some additional information based on true labeling( priority of classes to predict, minimum acceptable confidence, case complexity).  H-accuracy helps to curb the model drift towards 'over diagnosis'.  It can seen as a balanced, class and case weighted expression or measure of accuracy of a machine learning model  It is equivalent to 'Standardized Net Benefit' (very important measure to balance costs and benefits of diagnostic tools )measure (Kerr. et al, 2016)  The main novel contribution of H-accuracy is that it takes into consideration- the complexity of the cases of the test data set. This formulation of accuracy is safe w.r.t to the limitations that affect other common measures of model performance. |  |
|  | For H-accuracy implementation, annotation should set or code the following 3 types of information:  a. The threshold of minimum confidence the model should have to provide advice ( to penalize right prediction that have low confidence)  b.The priority of positive class w.r.t negative class( preference between sensitivity and specificity)  c. The complexity of each case in terms of rarity, difficulty, impact of missing, etc | ... |
|  | H-accuracy is completely backward compatible- In the formulation, if one parameter (TAU) is set to 50% then it is equivalent to regular accuracy. Then if 'P' (priority) parameter is set to 50% then it is equivalent to balanced accuracy and if 'complexity' parameter is set to constant, it is then it is equivalent to regular accuracy. |  |
|  | H-accuracy can be tailored to a specific diagnostic task by tuning the 3 parameters to make it more suitable to the preferences of the domain experts.The parameter configuration can be local ( e.g hospital setting) or for a specialist community, scientific society or association. That sense it can be considered as a parametric version of accuracy |  |
|  | Pi-Representativeness (Pi) :a simple and effective way to calculate the representativeness of a dataset (e.g. training dataset) w.r.t another dataset e.g. benchmark dataset representing reference population)  a.Pi-Representativeness is an alternative way of comparing data set distributions equality tests like Kolmogorov-Smirnov test  b. Pi-Representativeness helps us compute a measure of how similar is the training dataset to the benchmark test dataset |  |
|  | Robustness=(~Ha/Pi): A ratio of H-accuracy and Representativeness as a simple measure of Robustness and generasalibilitty of the model  a. If the training set and the benchmark test set are very similar, i.e. Pi=1, accuracy estimates are not reliable and hence then we cannot be sure that the model will have the same performance given a completely different data set  b.If the training set and the benchmark test set are significantly different, i.e. Pi=0, then accuracy estimates are reliable and that the moel skill will be maintained given new data set in real world conditions |  |
|  | Complexity' can be derived from the performance of the doctors based on the nature of mistakes and difficulty encountered during their practice. It can be also evaluated based on the doctors opinions with the help of Psychometric scales.  Here we introduce more of practical oriented 'pragmatic scale' , where the complexity of the case is evaluated according to 'which majority of doctors or professionals' should be able to get it right.  Pragmatic scale is based on the expectations of the domain experts on who should be supposed to diagnose the case rightly .  Right now the research effort is on how to combine different dimensions of case complexity to one single scoe- as a linear combination of different dimensions.the main dimensions are difficult, gravity (the model should be good at identifying serious cases) , rarity( doctors could be less good at identifying cases which they don't see often so model should be more accurate in detecting rare cases),urgency, etc. | How do you define case complexity? Can it be inferred from the properties f the dataset? What's the experience with the additional annotation burden needed for the computing the novel metrics? |
|  | ' Perceived Case Complexity' of doctors and practitioners needs to be given high emphasis and need to be taken into account while evaluating model accuracy validations.  The principle here is to make the AI models more robust in dealing with more complex cases of practice and H-accuracy as a measure is aimed on this principle.  The idea is to embed in benchmark dataset the multidimensional aspect of 'human perceived complexity'.  The advantage of additional effort of annotation on benchmark test set is that it could be done one time for every specific diagnosis task.  If AI model vendors or suppliers want to assess their models in terms of H-accuracy ,then they have to have their training set annotated in terms of complexity | How is the additional effort of labeling required in the proposed method to deal with case complexities in data sets justified ? |
|  | AI Model support in case of screening tasks (comparatively less complex) is well justified.  But at the same time, for diagnosis of more complex tasks, AI model needs to provide reliable and consistent support to the doctors. | The current paradigm may perhaps look at how AI models can better augment and help doctors in dealing with the ‘ simple cases(less complex) ‘ so that they can devote more time in dealing with complex case, where AI models still are not mature enough. How does it relate to the intended goals of the proposed method in helping doctors in dealing with more’ complex cases’ |
|  | The proposed metric have been, on an experimental basis,applied to the MRI dataset released by Stanford University. It was applied in 2 ways.  First one-the similarity between the training dataset and the test set was calculated to understand whether the model was made fit to data points that were close / similar to the datasets by which the model was tested. If the 2 datasets are very similar, it may be verified that the model is not being put to rigorous stress testing.  When there are is a new data point and if we want to understand whether data points are close to any other data point in the training set, you are assessing whether the response of the predictive model is reliable or not because if the new data is similar to any other point in the training set, it can be concluded that the model reliable recognises that new data point because of its similarity to the original data points on which the model was made fit.  So Pi-Representativeness can be used as a point-of-care interpretation tool to get a measure of robustness because here the accuracy score is normalized with a measure of how fair was the competition. | To what kinds of datasets that the Pi-Representativeness metric applied to? |
|  | Biases and Fairness are not considered. Here what is proposed is a measure of assessing the accuracy of a predictive model .  Pi-Representativeness is used to understand the extent to which the test data set is representative w.r.t the reference population and here in some way the common biases like gender bias , sampling bias are getting minimized because you are verifying that the datasets used are representative of the reference population | How does data bias affect the proposed metrics? |

Degrees of medical XAI

*Dr. Anne Schwerk, (Germany), H*

AI Health Project Manager - German Research Center for Artificial Intelligence (DFKI)

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Causality- cause and effect  Causibility- understanding of causal relationships  Interpretability- human understanding  Explainability- algorithmic/ model understanding  Global vs. local XAI  Global- how does the model reason  Local- individual decisions  Post-hoc XAI methods- problems= interaction effects, computationally intense, slow, lack of overlap between methods, lack of ground truth.  FairML is recommended as an end-to-end toolbox for auditing predictive models by quantifying relative significance of the model's inputs  Ante-hoc  Models e.g. decision trees  Criteria- simulatibility, decomposability, algorithmic transparency  XAI metrics  User satisfaction  Mental model  Task performance  Trust assessment  Correctability  Additional evaluation criteria for XIA  Confidence measures to training examples (representativeness)  Empirical evidence  Theoretical guarantees  Standards  Robustness, reliability measures, generalizability  Consistency of XAI models  Enforce experimentation to ensure validity  Ensure human-level understanding |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Explain the primary needs of XAI? | The 4 primary needs of XAI are  Control(AI should provide necessary control to improve system stability )  Liability (AI should provide explanation for their models internal logic such that the data and decisions could be traced and understood by the humans) ,  Generalization(Verification, Learning, enabling Replication and Transfer,Improvement, etc)  Fairness |  |
| Explain the Dimensions of XAI? | The different dimensions of XAI are  a. Explainability (Algorithmic or Model understanding)  b. Interpretability ( Human understanding of Features and Outputs)  c. Causability( Understanding of certain causal relationships)  d. Causality( Cause and Effect) |  |
| How can XAI Interpretability be classified? | Local XAI Vs Global XAI  Local XAI : Ability to understand individual decisions for a particular case or feature  Global XAI : Tries to explain-How does the model reason? |  |
| Explain the methods of Posthoc XAI ? | Posthoc XAI methods consists of :  a. Input perturbation LIME, SHAP method, occlusion, etc  b.Signal method-input method based on activated neurons (Activated Max.)  c.Proxy mechanism: simplifying ANNs, e.g. DeepRed |  |
| Explain some of the challenges of Posthoc XAI : | Challenges of Posthoc XAI :  a. Interaction effects  b. Computationally intensive  c. Slow  d. Lack of overlap between methods  e. Lack of Ground Truth |  |
| Explain the methods of Antehoc XAI ? | Antehoc Local XAI methods include:  a.Verbal decision path  b.Heuristic input attribution(Saabas)  c.Posthoc XAI |  |
| Explain some of the challenges of Antehoc XAI : | Challenges of Antehoc XAI :  a. Insufficient for multiple trees  b. Tree-depth bias for feature relevance  c. Slow and Sampling variability |  |
| Explain the Measures of Explanation Effectiveness for the DARPA XAI Psychology Explainable AI Program? | 1.User Satisfaction  a.clarity of the user explanation( user rating)  b.utility of the user explanation( user rating)  2.Mental Model  a. Understanding individual decisions  b. Understanding overall model  c. Strength /Weakness assessment  d. 'What will it do' prediction  e.'How do I intervene' prediction  3.Task Performance  a. Does the explanation improve user's decision,task performance  b. Artificial decision tasks introduced to diagnose the user's understanding  4.Truth Assessment  a. Appropriate future use and trust  5.Correctability  a.Identifying errors  b. Correcting errors,continuous training |  |
| Explain the typical challenges in different transparent AI Pipeline stages? | 1.Data collection  a.Sampling / labeling Bias  2.Data Wrangling  a.Standardized cleaning  b.Data Lineage  3.Feature Engineering  a.Overfitting  4.Algorithm Optimization  a.Efficiency based optimization or not  5.Model Training  a. Suitable outcome metrics or not  b.Model robustness  6.Cross-Validation  a.Representative data  b.Error analysis(any guarantees that errors or accuracies are equally distributed or not)  7.Deployment  a.User Satisfaction  b.Equal True Positives or False Positives or False Negatives |  |
| Explain the Degrees of Good AI Explanation Systems | 1.Content  a.Faithful to the model ground truth or not  b.Causal relationships, counterfactual faithfulness  c.Interacting Features  d.Limitations  e.Low Dimensional  f. Logical  2.Psychology  a.Useful(effective:trust,use)  b.Simple  c.Format : verbal, visual, etc  d.Clear, Precise, Complete  e.Granular  f.Example based  g.Flexible(end-user,context) |  |
| Explain the need for XAI Standardization | 1.No ground truth for Post-hoc XAI  2.No overlapping Post-hoc XAI methods  3.Unconsidered AI Pipeline  4.Unaddressed need for user-centred XAI(adaptive, interactive) |  |
| Explain the co-creation framework for interactive and explainable ML | 1.Understands ML models  2.Diagnose model limitations using different explainable AI methods  3. Refine and optimize the models  4.Stages involve-XAI strategies->Knowledge generation->Provenance tracking->Reporting and Trust building |  |
| Explain Evaluation Criteria for XAI | 1.Confidence measures to training examples  2.Empirical evidence  3.Theoretical guarantees  4.Standards such as IEEE -P7001-Transparency of Autonomous Systems  5.Robustness, Reliability measures, generalizability  6.Consistency of XAI models  7.Enforce experimentation to ensure validity  8.Ensure human-level understanding |  |
|  |  | One of the main underlying challenges is that the term 'Explanation' itself is not well defined and it is not defined for different ed users and different scenarios .  Also there is a lack of ground truth that can compare the performance of different methods of explainability used. |
|  | The main aim of 'explainability' should be to improve the 'usability' of the AI system  We can adopt a usability test protocol to evaluate the performance against 3 error conditions -Wrong Perception, Wrong Cognition, Wrong Action (PCA method ):  1. The possible risks are considered-What could happen if the user misunderstands recommendations by the system  2. This defines the requirements of Use Interface for safety relevant use scenarios  3. A usability test plan is prepared with the support of the particular representative of the user group to define the targets.  4. The targets conform to conditions that the end result is achieved completely without help and without mistakes. i.e. | How can 'explainability' be explained from a 'usability' under a regulatory consideration (looks at usability of explanations) ? Will the interaction with the 'explainability element' of the AI system be included in the usability evaluation? |
|  | ISO 9241-11 definition of usability is: “the extent to which a product can be used by specified users to achieve specific goals with effectiveness, efficiency and satisfaction in a specified context of use.”  Regulators typically do not evaluate 'user satisfaction' parameter directly unless it found to influence user safety and risk avoidance | Is user satisfaction taken into consideration in 'explainability'? |
| AI models with good performance have many parameters- difficult to have understandable explanation | Human agents in the loop | Causibility is also (only? See Holzinger 2019) related to the usability of explanations, and the quality of explanations with respect to their capability to trigger better understanding, user satisfaction, and trust (see the DARPA framework) in decision makers. |
| How to make the whole AI pipeline transparent?  Data collection  Data wrangling  Feature engineering  Algorithmic optimisation  Model training  (Cross)validation  Deployment | ... | ... |
| What constitutes a good explanation? | Content- faithfulness to model, causality, logic etc  Psychology- usefulness, simplicity, format etc. | ... |
| No ground truth for post-hoc XAI. | Methods to quantify explanation e.g. identifying most important parts of image? |  |
|  |  | Usability testing- with representatives of groups that will be using AI. Consider what needs to be achieved to ensure positive clinical outcome?  PCA model- what could go wrong?  Testing- can the users read the result? Why did users choose a certain option? Observe- what did the user actually do? |
|  |  | Methods of permitting feedback from doctors/users to manufacturers on usability, not just performance. |
| Uncertainty measures e.g. with Bayesian methods compare levels of certainty- again no benchmarking as no ground truth. |  |  |

## Data WS

Original file: [*https://docs.google.com/document/d/1v\_WSiGrFEByG1m34EWUe9prb9sA\_vJBBPsAdkZvjqdQ/edit?usp=sharing*](https://docs.google.com/document/d/1v_WSiGrFEByG1m34EWUe9prb9sA_vJBBPsAdkZvjqdQ/edit?usp=sharing)

Reference Data for Machine Learning

*Prof. Dr. Tobias Schaeffter, (Germany), H*

Head of Division, Professor - Physikalisch-Technische Bundesanstalt, TU Berlin and King's College London

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Division of Medical Physics and Metrological IT: aim is to develop new quantitative measurement techniques and reference methods for precision medicine. Provides mathematically sound approaches for data analysis and ensures data security in legal metrology (e.g., MR-tomography, biosignals, biomedical optics, modeling/data analysis, and metrological IT).  ML in medicine: classification, data analysis, denoising/resolution/artefact reduction, and inverse problems (reconstruction). Challenges: regulatory approval of ML difficult, EU-Medical Device Regulation treats software as medical devices requiring clinical studies/trials, trust in algorithms cannot be fully understood (sensitivity to data bias/uncertainty and mislabelled data).  NIST (equivalent of PTB in U.S.) is active defining standards in AI. PTB (funded by Ministry of Economy) is collaborating with them.  FDA approvals have been based on simple cases (e.g., denoising of radiological images; segmentation), where you can get immediate feedback from a healthcare practitioner. It is more difficult to test classification. Thus, intended use and data are closely related. We need to consider this when identifying use cases.  Aim of classification in medicine: selecting appropriate therapy through classification of patient groups (“personalized medicine”; consider example of blood samples that create different results, need to classify methods rather than disease/patient as part of metadata).  Inverse problem: what is the influence of an algorithm on the data (e.g., MRI)? If data have uncertainty, how is this propagated? The motivation: fast imaging approaches (parallel imaging and compressed sensing) and long reconstruction times. With deep learning, we can reduce artefacts. Can this be trained on limited data? Yes, even with reduced training data we could get robust results.  Benchmarking “Grand Challenges”: we put reference data online, enable people to download them and create algorithms, and we tested algorithms (speed, accuracy). Who created the annotations? How many people are needed (what is interobserver variability?)  Synthetic reference data: Can we use synthetic data to avoid issues with this variability? Can we input raw data and define simulation parameters to serve as groundtruth?  Use case: ECG analysis (300 \* 10^6 ECGs recorded annually) automated analysis for arrhythmia detection and heart rate variability (HRV) analysis. In a recent study, raw ECG data were viewed (rather than spectral data) for analysis to detect arrhythmia.  Use case: EMPIR project (Metrology of Automated Data Analysis for Cardiac Arrhythmia Management), we could choose model parameters, create an electrophysiological simulation, project these values on a thorax and create virtual measurements, and then produce a virtual ECG. The goal of the project is to create 5000 virtual patients.  Uncertainty analysis of ML and comparison to clinical experts (Clinical Turing Test).  PTB has clinical ECG reference database (22 \* 10^3 ECG recordings, 10s), 12-lead ECG measurements, diagnostic (62) and rhythm (24) statements according to ISO standard. Looked at superclasses and conduction disturbance.  Next steps for inverse problems: need to solve inverse problem with ML using biophysical framework and investigate physics-based learning. Can a vest be used to identify arrhythmic substrate in heart (instead of a catheter)? This is a growing market.    Conclusion: ML is of increasing importance for healthcare products, regulatory approval of ML is difficult, EU Medical Device Regulation treats software as medical devices requiring clinical studies and trials. Regulatory need for: digital reference data, benchmark tests, and definition of uncertainty. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing? (purple indicates a question from other participants)* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Sensitivity is mislabelled? | Consider data bias/uncertainty and mislabelled data |  |
| Where are ‘good’ use cases to find? | Consider classification where there is no ground truth (not just simple cases; e.g., denoising of radiological images; segmentation) |  |
| What to do when the same blood sample gives different results due to different methods? | Classify methods rather than disease/patient |  |
| Inverse problems- what is the influence of an algorithm on the data (e.g., MRI)? If data have uncertainty, how is this propagated? Even on limited data? | The motivation: fast imaging approaches (parallel imaging and compressed sensing) and long reconstruction times. With deep learning, we can reduce artefacts. Even with reduced training data we could get robust results. |  |
| Differences in interobserver annotations | for example, use ML for artefact reduction, reduce the amount of training data needed to reach robustness  Benchmarking contains grand challenges (100 different data sets, international initiative) |  |
| Can we also use synthetic data? | Use case: EMPIR project (Metrology of Automated Data Analysis for Cardiac Arrhythmia Management), we could choose model parameters, create an electrophysiological simulation, project these values on a thorax and create virtual measurements, and then produce a virtual ECG. The goal of the project is to create 5000 virtual patients.  -Synthetic ECG-Data with  Synthetic activation Maps, virtual measurements, electrical activity of the heart > creating an ECG, goal: 5000 patients with healthy and unhealthy person.  Benchmark test for better understanding of AI-tools  Comparison to clinical experts ‘Clinical Turing Test.’ Top 5 European cardiologists (senior) were identified (with a few junior for day-to-day work) and were easy to recruit. They are being paid for data.  Clinical ECG reference data sets should be published in ca. 6 month (in cooperation with Fraunhofer HHI)  Main goal of reference data set was to find a way to get a good distribution in training and testing data. Certain subcategories (at least 10) were created. There is also a free text field with various languages represented.  Use case: ECG analysis (300 \* 10^6 ECGs recorded annually) automated analysis for arrhythmia detection and heart rate variability (HRV) analysis. In a recent study, raw ECG data were viewed (rather than spectral data) for analysis to detect arrhythmia. | ... |
|  |  | Regulatory approval of ML is difficult |
|  |  | EU Medical Device Regulation treats software as medical devices requiring clinical studies and trials |
|  |  | Regulators require digital reference data, benchmark tests, and definition of uncertainty |
| Privacy of data is a problem -how can we deal with that? | Patients see data as ‘their’ value; the speakers impression: Patients are happy to share their data.  Create synthetic data base because data is needed now, make people aware to contribute data (like organs ) |  |
| How do we make data available to the community? What is the process? | For MRI, we are looking at society level; trying to get a data registry from various venders. However, defining standard on how to format the data is very difficult. The greater issue is rather finding an agreement among the various stakeholders (each company wants to have “the best” images, there is no incentive to harmonize on a fine level). To bring parties together, we first need to address the fact that everyone wants to have a leadership position. | We should care more about making models domain adaptive than data homogenization. |
| What are the challenges with anonymizing data? | Depends on the data (ECG easier to start with, detection weight is not as high as genetic data) |  |
| Can you generate anomalies (unhealthy patients) based on prior knowledge? How is that simulated? | Hope: people use software open source, and donate data back | Data donation is the best solution for the future. Consider blockchain use. However, we need to agree on one cryptosystem for compatibility. |
| How can we apply this information (about synthetic data) for creating evaluation platforms? How should WHO encourage hospitals/clinics to share data? | If you are a government-funded institute, you should be required to donate data. The only piece that is missing is the privacy requirement (for data sharing). However, this is not happening in reality. It should be enforced.  Maybe ‘stopping the grants’ to encourage people to make data accessible | This could be a practical problem (there are too much data). |
| How to ensure that data are correctly described (metadata) so that they can be used? |  | Can contact organizers of platforms. |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

The Long Tail of Medical Data

*Thijs Kooi PhD, (Germany), H*

Machine intelligence engineer - MX healthcare - https://www.varahealthcare.com/

Information about the topic: https://medium.com/merantix/the-long-tail-of-medical-data-fa31f6e9f9c

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| -Zipf’s law: long tail/asymptotic distribution of 100 most frequent words in Wikipedia.  -How to discriminate Japanese from English works using sample data. Can extract features (length of word in Roman  characters; number of syllables).  -Discriminative models: does an outlier end up on the correct side of the trend line?  -Power law in continuous space: what images are common or rare? This applies in medical imagery.  -Consider interactions among AI and medical images: detection air (computer provides markers), decision aid (computer provides a score), soft triaging (computer ranks cases), hard triaging (computer selects only suspicious cases), and autonomous (computer reads images without human; our ultimate goal). Kooi is working on hard triaging. For first three instances, humans are involved in the procedure. For the last two cases, computer functions alone.  -Consequences: computer aided diagnosis (doctor interacting with a system that misses obvious case could cause doctor to lose faith in system).  -Breast cancer: most common cancer among women (⅛ women will get in life), still one of the most lethal cancers in absolute numbers, early detection can reduce mortality by up to 60% if diagnosed in a timely manner. Net benefit of screening is a debate: false positives (lead to overdiagnosis) and false negatives (20%-30% cases are missed).  -Computer-aided detection: several systems are FDA approved and 91% of digital facilities in the U.S. use CAD. Effective for calcifications but for masses is less clear. NEJM and JAMA papers have highlighted weaknesses.  -Smart rule-out: instead of proving CAD markers, trying to filter out successes.  -Use case -- Late stage breast cancer: women are screened each 1-3 yrs; if you skip a screening, cancer can reach late stage.  -Use case -- Inflammatory breast cancer: rare but aggressive (ca. 1% of breast cancer cases). In database of ca. 2 \* 10^6 images, ca. 20 samples.  -Use case -- Phyllode tumors: rare but benign carcinoma-sarcoma hybrid (ca. 1% of breast cancer cases).  -ML discriminative model (directly estimates probability and creates decision boundary; e.g., regression) versus generative model (estimates probability of something and uses that to infer probability of something else; creates probability distribution of data).  -Testing: metrics of AI studies do not take anomalies into account. Furthermore same performance does not mean same behavior.  -Solution: creating an anomaly database.  -Current ML solutions in radiology only use discriminative models. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Learning Problem:  -Distinguish Japanese from English words, or images in general vs.  medical images | Extract two features (Length of a word and syllable)  Discriminative models  Detection aid (Computer provides markers) -> Human checks  Decision aid (Computer provides a score) -> Human checks  Soft triaging -> Human checks  Hard triaging (only suspicious cases, that’s what MX health care is working on)-> ML Generative model, discriminating model  Autonomous | [YOUR NOTES HERE] |
| Net benefit of screening is a debate: false positives (lead to overdiagnosis) and false negatives (20%-30% cases are missed). |  |  |
| When testing AI - metrics of AI studies do not take anomalies into account. Furthermore same performance does not mean same behavior. | Creating anomaly database. |  |
| For cases where there is a lack of data would synthetic data be useful? | We do not know the full range of anomalies that can occur. Therefore, things that are rarer might not be represented in our dataset and therefore cannot be simulated. | Create more data in case where there is a lack of examples |
| Why is an abnormal sample / image obvious for humans and not for AI? | Humans build a model of the world whereas models just learn to discriminate things (generative model versus discriminatory model). |  |
| What methods have you tried and what have failed? | We tried simple methods, can’t provide details. Mostly looked at simple outlier detection methods (deep learning and classical). | Cluster analysis is an issue because you are moving high-dimensional information into a lower-dimensional space. |
| Among the anomalous data, do you have some “normal” data? Sometimes, you don’t know what is normal and that makes it difficult to do outlier detection. | Because we are looking at both false negatives and (occasionally) false positives, we do manual identification. |  |
| Have you tried meta learning approach (training models similar to how humans perceive things)? Could you train model in a similar way? | Interesting academic approach but not practiced at their company. There are some people doing Bayesian methods but we are not. |  |
| Is it possible to make the anomaly dataset open? | No. That’s the value of the startup: data curation. |  |
| What is accuracy on filtered out data (once 40% of cases are removed because they are inaccurate)? | For a specific type of cancer, it can be up to 98%. |  |
| For future versions of your software, how will you handle anomalies? | Further research on this has to be done. |  |
| Do you explicitly tell users (radiologists) the cancers for which your software is not as good? | Yes. |  |
|  |  |  |
| ... | ... | ... |
| ... | ... | ... |

Towards Deep Federated Learning in Healthcare

*Shadi Albarqouni PhD, (Germany/Switzerland), H*

Senior Research Scientist - TU Munich and ETH Zurich

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Medical image Analysis TU Munich  -54% of healthcare leaders see an expanding role of AI in medical decision support. Can create gains but also cause pain.  -Deep learning has been shown to offer solutions that outperform average clinicians in various fields.  -Most papers are published in Nature and buoyed by considerable amounts of data.  -Technical challenges data volumes, data quality, data privacy, model robustness/fairness/explainability, and user uncertainty  -Federated Learning: distributed ML approach that enables training on a large corpus of decentralized data on devices such as mobile phones. Differs from distributed learning because data are not identically distributed and it’s communication efficient. Issues: expensive communication (model compression; subsampling, quantization), systems heterogeneity (active node sampling, fairness/accountability/interpretability), data heterogeneity (highly imbalanced data, intra/inter-scanner variability, intra/inter-observer variability)  -Research: learning to recognize, adapt, learn, and learn from past knowledge, learn to reason and example, and then federated learning.  Learn to recognize (Semi/Weakly Supervised learning, classification task, e.g. in MRI (MS-Lesions)  Learn to adapt (segmentation of lesions on different MRI from different companies; SainNormalization), adversarial Networks for Camera Regression and Refinement;  Learn to Learn: Semi-supervised Few shot Learning (provide model with CT-Scan, and support set (e.g Image kidney from another patient), provide model also with Bounding boxes; AI model only has ‘one shot’ > conclusion: the AI model did a good job  Learn to reason and explain: Train model in robustness, while training doing classifications, explain why a diagnostic decision is made; Learning interpretable Disentangled Representations; Bayesian Formulation, Objective function, Labels uncertainty Estimation, how physician would agree on decision (e.g. 90% of cardiomegaly of system> double check with physician)  Federated learning: Data Heterogeneity  Can we detect and diagnose those heterogeneity? How to handle such heterogeneity?  Conclusion:  Generalizability: shown that models were generalisable, adaptable  Robustness: shown, that models were robust, but Direct quantify still missing  Uncertainty Quantification  Explainability |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Data volumes (need of a lot of data, heterogeneity, missing data, inter-scanner variability, noisy annotations, bias), data quality, data privacy |  |  |
| Model robustness, fairness, explainability |  |  |
| How to address user uncertainty and provide quality control? |  |  |
| Federated Learning: Problem Formulation  Key Differences with Distributed learning, concentrate on Data Heterogeneity | Learn to recognize (generative models, PGAN-Models, segment Anomaly), to adapt, to learn from prior knowledge , learn to reason and example |  |
| Q: Infrastructure |  | Hospitals are willing to cooperate |
| Q: What is the trade of directly training with respect for doing mapping data? | Data needed from 2 different domains,  Transfer the knowledge |  |
| Q: How do you handle the problem with models that have been trained on e.g. MRI from different companies ( e.g. Siemens or Philips ) | Different hospitals use different scanners > ‘model poisoning’ (problem, that you train model on bad annotations) > How can we detect a ‘poisoned model’ before the update? > Future thoughts on that issue |  |
|  |  | Cluster federated learning  In the beginning all learn together, between the updates the models learn to specialize in clusters  Only a part of a model has to be updated  Sometimes it doesn’t make sense to have one model for all |
| Q: How to make use of federated learning? | Different network architecture; hardware different, technology changes quite fast |  |
| Q: From a practical point of view: How do you acquire data? | Work at the hospital, generate data locally |  |
| Q: What about models that are forgetting data? | Federated learning should overcome this problem, long tail examples | Models learn not from complete data source because parts are forgotten |
| Q: Do you train on global data or regional hospital data? | Speaker hasn’t run in such a problem | Challenge to deal with privacy issues in this context |
|  |  | From Slides: C2 and C3 the most important issues are Systems and Data Heterogeneity |
|  |  |  |
|  |  |  |

An Alternative to Federated Learning: Generative Adversarial Networks for Data Sharing and Data Privacy

*Auss Abbood, (Germany), H*

Research Scientist - Robert Koch Institute

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Speaker is part of the TG Outbreaks, with this also interested in Privacy issue, generative adversarial networks of data sharing and data privacy  -General motivation in TG: benchmarking important to put new methods to the test, especially data is rarely publicly available, realistic data is hard to get; so models like federated learning are interesting, to bring AI to the data  *Pros* for FL (Training of models runs locally, enables access to more data, several small models can be trained, packages help with implementation’)  *Cons:* Computation depends on local device, no platform for development (cf. Spark), No validation possible during training (e.g. to control overfitting), Local data is not identically and i.i.d., might require preprocessing/hyperparam. Tuning onsite  > Handle cons:  Anonymized data can compensate the cons  > bring data to one place  Anonymization  1-to-1 transformation of original data  Explicit probabilistic model  NEW: Generated data , produced by GANs  How to generative adversarial networks  How Neural networks work: several hidden units; auto encoder (takes data and tries to compress data to do decoding)  GANs: True data distribution (e.g. sample, random noise) Generator and Discriminator (fake vs real)  Random noise > Generator > training set > Discriminator ;  Paper was inspiration: https://arxiv.org, usable for topic group  Experiment:  GAN trained on reported case of Salmonellosis in Germany: Generated vs. real data > Results: Real data shows the same effect in generated data; average case count per week of generated data  Conclusion:  GANs can create data that can be analyzed in a central way  Data from different sources simultaneously (e.g. from different institutions)  Data becomes shareable  Experiments become reproducible/comparable though they are based on sensitive data  Outlook:  GANs are hard to train  Hard to evaluate  Loss of information is hard to estimate  Q: Are GANs (for certain domains) a ‘better’ alternative? |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Are GANs (for certain domains) a ‘better’ alternative to federated learning? | For discussion |  |
| Q: When we think of benchmarking (e.g. in Focus group AI4H)  Is there still mismatch in gaining trust in clinical setting for GANs? | Outbreak detection is written on already trusted data |  |
| Q: GANs Alternative to federated learning - how do you get access to data? | Test a GAN on data from platform of the RKI, publish it as new ‘fake data’ for public use in a data - > data pool, find maybe clusters |  |
| Q: Encoder / Decoder? | Train outer encoder forehand > the performance increases | Outer encoder is memorisable (it’s not private anymore) |
| Q: Privacy of GANs data | (considering the paper): generated data was tried to re-identify people > very small amount could be identified  There are Companies (e.g. Startup in Berlin) working on this question | Idea of sharing fake data instead of real data  There should be further considerations about guaranty of Privacy of data in using GANs |
| Q: How does the model collapse with the data? | Created time series over one year;  You don’t need more than 5 years for time series |  |
|  |  | Medical imaging: few amount of data; building new models with this small amount of data is difficult  FL feed more data (like weather, social), difficult for AI |
|  |  | Outbreak data (time series) labels for the real data is requested; people that report to the RKI ; ground truth problem (that’s why the RKI doesn’t use labels so far) |
|  |  |  |

Diagnosis of Brain Diseases with AI Applications in Federated Data Sharing Environment

*Ferath Kherif PhD, (Switzerland), H*

Vice-Directeur - Laboratoire de Recherche en Neuroimagerie (LREN), Université de Lausanne

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Data sharing in a federated environment  Topic Group - Neurocognitive disorders (TG-Cogni)  Use case: BrainLaus (cardiovascular disease?): advancing fundamental and digital knowledge on healthy brain aging and neurocognitive disorders (establish a framework for federating clinical data directly from hospitals, developing a benchmarking technology, evaluating AI-based diagnostic, and deriving biological signatures of brain disease).  Medical intelligence: framework enables collection of data without centralizing the data (“Data Integration Platform”). Researcher can login and ask questions. Computation is done locally. Each hospital has a server where data can be captured, preprocessed, and ML can be locally conducted.  Mission: to create collaborative network between different hospitals/clinicians and researchers. Allows sharing of medical aggregated data and use of advanced ML modeling and statistical tools for brain-related diseases while preserving patient confidentiality.  Vision: break down traditional barriers between patient care, brain science, and clinical research.  Multiscale approach: from cell to system level (focuses on complete pathway of each brain disease; identifying biomarkers from behavioral to proteomics and brain imaging to get “biological signature of disease”). Various features provide a broader view (with physical and causal features combined with clinical features).  Biomarker evaluation: deep phenotyping, stratification, diagnosis.  Clinical validation: there are some measurements (e.g., MRI) that have high variability, which makes it difficult to use repeat measures. In federated approach, you can test robustness of measurements.  Clinical utility: can be viewed from two perspectives.  Mathematical: difference in accuracy between standard method and new (AI) method.  Clinical treatment: based on results of AI (diagnostic), clinical treatment will be modified. Is there a difference between the system and the opinion of the clinician?  Use case: used PET, MRI, genetic data, CSF, and protein data to look at the difference between Alzheimer’s patients and healthy patients. Several challenges: not complete data and more data than patients. Thus, needed to adapt methods. For first proof-of-concept, found that there were various clusters that did not clearly match the expected characteristics.  Error rate of clinicians is 20%-30% for diagnosis of Alzheimer’s. We want to use post-mortem validation data of people who were symptom free but diagnosed with Alzheimer’s.  Goal is for early diagnosis of people who are still symptom free.  Lessons learned: technical (infrastructure) data challenges (combine multi-modal, multi-scale data, personal health sensitive data from multi locations), data management and governance solutions (federated platform with built-in data privacy by design, decentralized processing, and distributed analysis).  Organizational challenges: installation (customized configuration for each location hospital), data sharing (privacy and heterogeneous data), and point-to-point communication protocols  Infrastructure solutions: customized cloud deployment, data as a service (data is accessible to authorized user), semantic interoperability, and standard data communication protocols. |
|

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| How can we improve our understanding of processes at cellular level in a special disease (e.g. Alzheimer’s disease) that can’t be measured so far? | Biomarkers validation, deep phenotyping, Stratification, Diagnostics, collect data in an accurate way |  |
| How do you handle environmental varieties (e..g different MRIs in different clinics)? | Observe minimum requirements in federated data sharing |  |
| Can AI tools measure more accurate than humans? | In some aspects, yes |  |
| Error rate of the clinician is 20-30%, how do you measure? | We want to use post-mortem validation data of people who were symptom free but diagnosed with Alzheimer’s. |  |
|  |  | Different types of data  Federated learning  Customized configuration for each location hospital, Data sharing |
| Organizational challenges (installation, data sharing, and point-to-point communication protocols) | Customized cloud deployment, data as a service, semantic interoperability, and standard data communication protocols |  |
| Do patients know that they are part of the research? | At the involved hospital, there was implicit (opt-out option) consent. Now it is switched, patients need to opt-in. However, the patients should be aware that they do not need to be part of all kinds of research. This is a challenge especially because the data are anonymized. |  |
| Are data saved and stored in server? | Data flows from hospital center database, then anonymized, and brought to a local database. These data are then accessed. Each node, however, can create it’s own network. |  |
|  |  |  |

A Federated Convolutional Denoising Autoencoder for MRI Applications

*Dr. Alberto Merola, (Germany), H*

Machine Learning Scientist - aicura medical GmbH

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| AICURA’s vision for federated learning: edge computing platform uses federated learning. The data remain in the hospital and can be used for ML application development. In addition, hardware is provided (GPU power for AI computing) with real-time integration into clinical decision processes and local IoT.  Virtual data pool should be accessible for training purposes from third-party developers/researchers.  Convolutional denoising autoencoders (CDAE): state-of-the-art method for MRI image denoising, a crucial step towards image quality improvement. Performance depends on training data with noise features as varied as in reality. This is done via aggregating big data sets from different sources with varied noise characteristics.  There are practical limitations (aggregating data from different hospitals into a single database) and privacy issues (pooling data off premise).  Solution is federated convolutional denoising autoencoder: enables training on small datasets from different hospital slocally and aggregating training models.  Method: data divided into four virtual hospitals (ground truth) and noise was added (thermal, zipper, Gibbs ringing, k-space spikes, and ghosting). For each virtual hospital, a unique combination of the tuning parameters for artifacts was used to simulate real-world differences in noise patterns between hospitals. CDAE architecture has encoder path and decoder path. Each path has three steps (convolutional layer, batch normalization, and leaky ReLU). Fully connected layer was included for high noise.  Training was layer-wise and weights were optimized with ADAM. Loss function was Negative SSIM.  Qualitative results: for low noise levels (original vs noisy MRI), gradient of quality is apparent. For high noise levels, single virtual hospital results are not useful. In federated case, most important structures are already apparent.  Performance quantification (SSIM) showed improved performance in pooled CDAE. However, FCDAE outperformed single virtual hospital case and, in particular, is robust against outliers. In clinical setting, this is important because we want robustness rather than excess variability.  Conclusion of performance:central combination of locally trained model weights results in more robust and better performing denoising solution, compared to CDAEs trained separately.  FCDAEs solve practical and privacy issues, allowing:  Training on small(er) datasets with no need for data aggregation  Effective models combination for handling high noise heterogeneity  Local and privacy compliant medical data analysis  FCDAEs is a simple solution for dynamic and scalable aggregation of ever more datasets for closing the performance gap with CDAEs trained on centralized datasets. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| How can the hospitals handle the AI tools if the technical standards are different? | Provision of hardware (with GPU) and virtual data pool |  |
| How can we improve the quality of MRI? | Convolutional denoising autoencoders (aggregating big data sets) > leads to more robust and better performing denoising solution |  |
| How do you handle practical limitation  and privacy issues? | Train model locally on smaller data in the hospital; data from an open data set, split those in 4 virtual hospitals plus add noise (real life artifacts, like zipper artefacts, Gibbs ringing); training artefacts were different in between the different hospitals > *Architecture* Convolutional Layer, Batch NOrmalization, Leaky ReLU; *Training Algorithm* |  |
| How do you approach hospitals to set up infrastructure for federated learning? | First, we install our system within the existing IT landscape (firewall) and give permission for remote communication. It is important that our communication goes through the existing IT firewall. Thus, IT is aware of the data flow. |  |
| Ransomware hack affected hospitals in the past. Thus, you probably need an administrator to explain the IT system. Have you had any issues with staff? | It is a collaborative solution between AICURA and hospital; adapting to their needs. However, then the platform takes care of all the overhead. |  |
| Risk for diagnostics with denoising - will AI interpret a lesion as noise? | Not easy to answer.  Denoising for pathological data is planned in future |  |
| How do you approach the hospitals to get access to data? | Depends on individual doctor you are talking to (different interests) focus on different use-cases, e.g. for research or accounting data (the bigger the hospital the better),using the structure that exists in the hospital |  |
|  |  |  |
|  |  |  |

Validating Image Compression for AI in Microscopy

*Dr. Bruno Sanguinetti (Switzerland), H*

CTO - Dotphoton AG - https://dotphoton.com/

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Validating image compression for AI in microscopy (growth of image data in health, raw vs compressed data, validating compressed raw data).  Digital image versus viewing directly through microscope: difference in depth and detail.  More data is digitized and kept (required from scientific journals, traceability, reduction of animal testing, training AI, point-of-care diagnostics with remote processing, and scaling academic research to clinical use).  Data volumes increase faster than disk capacity: histopathology (4 TB per patient), lightsheet microscopy (up to 20 TB per image), high-throughput screening (10 GPS; 10 TB per hour), and GPU.  Data retention costs (semi subsidized): in CH (10 yr, 3 backups, ca. 3 K CHF per TB); in DE (30 yr, 7 backups, ca. 10 K CHF per TB).  Image data need to be: adapted for computer vision, interoperable, fast to store/transmit/process  Typical image pipeline: irreversible, implementation dependent, adapted to human eye.  There is a sensor, an image is digitized (there is some correction for pixels), there is scaling (... processing prepares image for eye …), there is a compressed image. This pipeline is nonlinear and can vary according to user. However, AI usually functions better on raw data (rather than compressed imagery). The raw data are standardized because the photons hit a sensor creating an integer which serves as a standard.  AI issues with visually lossless compression: biases (amplitude: rounding errors, quantization; spatial errors: loss of detail and artefact creation; temporal), statistical properties are modified, and interaction between compression algorithms.  Consequences for AI: low interoperability (data acquired from different devices, data processed through different pipelines, difficult instrument/analysis pairing), require vast amounts of training data (difficult to handle for small institutions, difficult to diagnose rare diseases), low scalability (difficult to deploy results from one institution to another, difficult to roll out cloud diagnostics solution), and lock-in (manufacturer lock in, image format lock in, and uncertain future reusability of data).  Solution: machine vision lossless (up to 10:1 raw data compression). Loss is introduced only once at the beginning and loss must faithfully preserve image’s statistical properties. We introduce losses within the camera. Thus, raw data remain statistically consistent with the expected output of the camera. Afterwards, everything is lossless (you can convert the format of the image).  Validation: defining testing target (compressed images should be drawn from the same statistical distribution as the original images; there should be an accurate physical model of the image acquisition process; should take many images of the same subject to measure the statistics). Four steps: theoretical validation of algorithm, code testing (unit, integration, and system), laboratory testing (black box testing and visual element testing), and application testing (many acquisitions systems: light sheet, phase contrast, light field, fluorescence, spinning disc, optical tomography; many samples: test targets and real samples; many measured properties and methods).  For laboratory testing, each instrument model is tested. All levels of light are considered and we can correct all errors of a given sensor. Then, we can measure the statistics of the sensor.  Test scene is composed of technical and non-technical elements.  Shows example issues: error, bias, artefacts  Application testing: measuring, segmenting, etc.  If there is a strange result, they use explainability methods to identify cause. Typically, error was because the original data were modified without mention.  Normalization: two scanners are used that follow two preprocessing pipelines. Using a transform, we can try to reduce sample space (dof) so that AI trains more easily.  Data augmentation and error propagation  Additional features: per-pixel calibration, online image quality validation, authenticity and integrity validation, and metadata embedding. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| How can the quality of images be improved? | idea: if you approve high quality data then the AI delivers high quality as output; digitalisation and saving of more raw data |  |
| Why should data be preserved? | required from scientific journals, traceability, reduction of animal testing, training AI, point-of-care diagnostics with remote processing, and scaling academic research to clinical use |  |
| What are challenges related to data retention? | High cost; partly subsidized |  |
| How to deal with the challenges related to a variable compression pipeline? | Just use raw data for AI. |  |
| How does the AI deal with ‘visually lossless’ compression? | This is not targeted at post-processing which means that losses are not quantitatively specified. |  |
| How to design compression method for machine vision? | Possible solution: ‘’Machine-vision-losses’’ up to 10:1 raw data compression, Loss introduced only once at the beginning ; loss must faithfully preserve the image’s statistical properties |  |
| How to deal with validation? | Compressed images should be ‘’drawn’’ from the same statistical distribution as the original images  4-steps to validation (Theoretical validation of algorithm, code testing, laboratory testing, application testing) |  |
| What is raw data important for? | Keeping raw data for AI- quality,- validation,- interoperability, - error propagation |  |
| Data heterogeneity | Normalization of data with statistical model of instrument |  |
| Camera: linear and non-linear transformation- what happens within the transformation? | Apply early on camera |  |
| Does it happen that the AI saw a difference in an image that humans didn’t see? | Yes |  |
| What do you mean by validation? | Make sure that algorithms would maintain data |  |
|  |  |  |

## Assessment Platform WS

Original file: [*https://docs.google.com/document/d/1kIAF9NqT6v0Z-JpJb\_jDZUp9ub6ZtqySTb7TYv\_VdMQ/edit?usp=sharing*](https://docs.google.com/document/d/1kIAF9NqT6v0Z-JpJb_jDZUp9ub6ZtqySTb7TYv_VdMQ/edit?usp=sharing)

Precision Evaluation: A Standardized Domain-Agnostic Framework and Platform For Evaluating AI Systems

*Darlington Akogo (Ghana), H*

Founder, CEO, Director of AI - minoHealth AI Labs/GUDRA - https://www.gudra-studio.com/

Precision Evaluation presentation

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Tells you how your AI solution is performing over different populations. Offering and open-sourced prototype for the FG |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| We have systems developed and assess their accuracy AUC for instance. Cannot guarantee the numbers published. | Agnostic benchmarking platform and framework. | Worked on a similar platform. Did not solve issues: Black box evaluation, models encrypted. Keep the evaluation data private. How to find data which are very expensive ? Who is going to annotate the data ?  Data donation incentive. How well is this AI solution working for your population. |
| We need standards | Best suited for classification. | ... |
| ... | AI system would be compared with domain experts. | ... |
|  | Evaluation divided into Location, Gender and Age. We can add subcategories for Location: Country, Continent, Region and Global. Ability to generalize or expand to a broader region. | Cannot collect unbiased data. You will always get skewed data. Variance issues |
|  | Evaluation data has to be diverse, coming from different populations across the world. |  |
|  | Panel of domain expert to analyse the submitted data. |  |
|  | Test human experts. From different parts of the world. Compare German doctors with German patient data and not US. | Why have a solution not working in India as well as in the US ?  Tells you what is the best solution for your context, very precise case and not the best solution globally. Not a Kaggle leaderboard. |
|  | Privacy and security. Abstracting date of birth to age. Eliminate risk of linkage attacks because dev and org do not have demographic data. Dev and org needs to be registered before having access. At a local WHO or ITU branch. In-person registration. Pay a fee. |  |
|  | Don’t have to submit the model itself. Black box submission. No IP problem. |  |

AWS Sagemaker + CI/CD + Docker as a Platform Blueprint

*Dr. Steffen Vogler, (Germany), H*

Senior Data Scientist - Bayer Business Services GmbH

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Use a well maintained Container Registry with a CI/CD pipeline to build flexible and reproducible assessment platform. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Used AWS to develop computer vision (CV) solutions | Use amazon Sagemaker: cloud based service for data labeling, code env and deployment. SDK for Python | How this setup can be useful for testing ?  Use a template to publish your web service. |
| Developed CV workbench | Sagemaker “introduces” Env-standardization. AWS pulls a standardized docker image from its repository according to the parameters value. On top of the image add your own scripts | Possible to do it without Cloud access. AWS service is not required. There is an open-source version of every component. |
| In ML everything takes longer than software dev. ML agile is not going at the same speed as dev agile. Need a standardized framework. DS need a sandbox | You can build your own custom image from a base image or from scratch. Can add different versions of your image and build an archive of containers. In terms of reproducibility you can go back to your previous image. | Submit a Docker image with XX port opened and Rest API that receives images and sends back the result. Any limitation ? No |
| Being able to reproduce model training and testing. Version control on code and data and hardware (accuracy limited by computing capacity). Monitor and log environnements | CI/CD setup feature from Gitlab (which is open source and can be deployed on premise). A way to bring new features to the customer. Check for bugs, errors before launch. This is highly automated. Continuous Integration and Deployment.  For ML evaluation we could use this pipeline. | Can it be a black box model ?  It is language agnostic. Needs to publish predefined endpoints. |
|  |  | This solution is good if you want to deploy and maintain an AI solution. Is it also good for a benchmarking platform ?  What if you cannot pack your AI solution into a Docker image ?  Organize benchmarking sessions at the same time for all participants. Then you can use only one time your evaluation data. |
|  |  | What will be your recommendation for our benchmarking platform?  Use Docker and private Gitlab.  Can provide mockup, templates,... |

Clinical Validation Techniques for Medical Imaging AI - A Platform Approach

*Vasantha Kumar Venugopal MD, (India), R*

Imaging Lead - Center for Advanced Research in Imaging, Neuroscience & Genomics (CARING)

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Validation methods for different types of radiology AI solutions |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Validation methodology from a clinical viewpoint | Retrospective testing with simulated prospective arm. We are testing new design now. | Is your key recommendation to have different type of validation process for each type of AI solutions ? How to deal with “blackbox” submission ?  Yes, and we published a paper about that.  https://www.academicradiology.org/article/S1076-6332(19)30448-9/fulltext |
| Outcome analysis instead of numbers, improvement in care delivery. Improvement in efficiency. We have not reached this stage with AI Solutions.  Only retrospective testing. | Recommendations for validation studies. (see slides) | Is there an online platform available for testing ?  Yes, we can provide. |
| Broad category of radiology solutions: classification, segmentation, image translation,super-resolution. Each need a different kind of validation. | Testing different types of algorithms with specific methods | How to do you qualify validation experts ?  Many domain experts are coming forward to do validation.  How to get the best experts on the panel ? Experts clear experts |
|  | Algorithm audit framework and guidelines proposed  https://www.academicradiology.org/article/S1076-6332(19)30435-0/fulltext | How could you contribute to our FG ?  Documentation, research group. |
|  | Packaged into an assessment platform | Interesting to add metadata to FP and FN for further improvements |

A Tale on Data Quality

*Diogo Telmo de Sá Lima Pinto Neves, (Germany), H*

Researcher - German Research Center for Artificial Intelligence (DFKI)

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Data quality and impact of poor data quality on ML models |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Completeness, consistency, conformity, accuracy integrity and timing affects data quality. | Improve quality of the data encoding domain knowledge. | How can we use this tool for the benchmarking platform?  When third-parties send data to FG we could assess the data quality automatically.  We need tools to assess data distribution, dimension coverage, sanity check to confirm best practices have been used... |
| More is not always better | Purify library allows to assess data quality and clean data |  |
| Data errors: different types (error type taxonomy) |  |  |
| Today focus on missing values |  |  |
| Technics to complete missing values |  |  |
| What is the problem of missing values ? How can we cope with missing values ? Use DS to get rid of missing values |  |  |
| Model selection is feature selection, algorithm selection and parameter tuning. |  |  |
| Use Pandas profiling tools… Are we good ? Not really  Ex: Pima Indians Dataset  What is missing ? |  |  |

Big Data Governance

*Shobha Iyer (Germany/India), H*

Big Data Solutions Architect Consultant

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Data is a big asset of your company but in order to be able to use it effectively and efficiently you need to take care of it and put processes in place. They built a tool to manage Big Data. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Why do we need governance ?  No data transparency  Security risk  Multiple data silos | Data Governance Processes gives you: data tracking, data management, enforcing ownership, security policies, data catalog and integrity | How do you see the benefits of this complex architecture with regard to our initiative ? Regulators might be afraid of this complexity and you will spend months documenting the pipeline.  More useful from an application point of view and developer point of view. |
| How do we track, manage data ? | Master data, reduced duplicates, redundancy (improves data quality crucial to build ML models) | Do you have a template you could give to a regulatory body ? |
| We don’t want a Data Swamp | You should implement all components of the governance framework to get the mentioned benefits | Track which data has been used to test. Or aggregate datasets to build. |
| How to get values out of your data | Big Data governance tools : Apache Atlas, OMRS, Data Hub and Egeria  Implementation details in slides | The FG organization itself would benefit from using this implementation.  Some parts may be shared with the FG |
| What is data gov : formal process that enables sharing trusted data asset |  | The benchmarking system creates a lot of data that needs to be managed. We may need to go back to it. (Update the deliverables document to reflect this) |
| Challenges: Communication, ownership, shared agreement, management |  |  |

Enterprise Data Pipelines

*Dominik Schneider, (Germany), H*

Strategic Engineering & Principal Solution Architect in Advanced Analytics - Merck KGaA

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Re-use of data needs to be ensured, catalog should reflect data assets based on metadata, quality assurance of software artefacts is key. |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Provide curated data to generate insights most of your time instead of doing cleaning and curating. | Created MCloud engineering practices.  DataOps concept | For how long do you have this in place ?  Started 2.5 years ago, 1 year live. |
| Data offense vs Data defence  Focus on the latter: data governance, reusable, catalogued and quality assurance | Data Pipeline: staging, preparing the data and make it usable and standardized for projects to build a product.  Use Apache Atlas to track what was done with the data. | Do you have any feedback on what data-reuse changed ?  Maintenance made easier |
| Every project team build different structures, quality, code,... | Ensure to have good metadata from business | What are the implications for the data platform, benchmarking platform.  Used to track dataset combination, prepared in a controlled way. |
|  | Catalog: versioned metadata information about the dataset | Should this policy built top-down or bottom-up.  Top-down need to be done if centralized architecture. Need governance that gives guidance |
|  | Quality assurance: version control, branching. Can be used by the benchmarking platform to prepare the datasets |  |

AIcrowd Platform

*Sharada Mohanty, (Switzerland), R*

CEO - AIcrowd - https://www.aicrowd.com/

|  |
| --- |
| **Miscellaneous notes**  *Points you deem important which do not fit the below problem-solution-comment structure* |
| Evaluation platform |

|  |  |  |
| --- | --- | --- |
| **Problem raised in presentation**  *What challenges or problems is the speaker addressing?* | **Solution proposed** *What solutions is the speaker proposing?* | **Comments from other participants** What are other participants saying about the problems or proposed solutions |
| Benchmarking AI solutions | Accept code submission instead of a single CSV file with responses | How does the security aspect works ?  We suppose that all submitted code is malicious and take all measures to mitigate the risk (sandbox, no network access, ...). Rely on best practices and solution vendors ability to secure |
|  | Diverse and specific evaluators available | Do you have any requirements for input and output ?  Up to each challenge to define in out, in simplest case file-level; we require the user to provide a single entry point. |
|  | Benchmarking (FG - Snake bite) | How much is open-source ?  All but orchestrator |
|  | Submitted code is packaged in Docker image and orchestrated on Kubernetes Cluster | How much work is needed for a new challenge ?  1 week of work |
|  | Users can choose their solution to be open-source or not | AIcrowd allows the whole code to be shared with FG members. The backend ("Sourcerer") is not open source yet, but it will be made open source later in the year, under a non-commercial license. |
|  |  | TG-Symptom is building its own system to move fast and define the scores… and eventually to have a unified platform might come back to AIcrowd |
|  |  | Versioning of images and data is implemented in AIcrowd |

Summary

AIcrowd has a working platform, is willing to share.

Working:

Submission of algorithms in code

Easy management of different “challenges”, automated versioning

All code is malicious, safe environment, incl. No network access. Used to secret data.

Scaling of dockers to on-prem Kubernetes cluster

Metrics “leaderboard”

A few things that are missing

Some companies might not be willing to share code (IP)

Dimensions/”subgroups” dashboard

Dynamic metrics (“add new metrics” later, as required)

Timing of inference *per case*

Ease of changing / open source

Data donation & labeling & quality control

Carpl.ai : impressive dashboard backend, with different metrics and dimensions. Available for testing / open access.

Every problem/algorithm might need its own specific evaluation metrics, dimensions, dashboards

Dimensions (“subgroups”) to evaluate over: age, gender, location [, imaging vendor, ….]. Fairness across these?

These dimensions need to be collected with data from beginning

Baselines should be local to be realistic: developing world might profit from algorithms even if they cannot beat best-in-class radiologists

Data quality is relevant, and it’s tedious to achieve - specialized libraries can help, but manual work will always be needed.

Annotators: gold standard difficult to select, “experts name experts”

Annotation might need to be contextual/local to capture local specialties

Data governance: track provenance, processes. Apache Atlas has been successfully used elsewhere, is open source.

Relevant work goes into structuring the organization, possibly more than tech

Enterprise Data Pipeline: Good solutions how to avoiding data swamps.

Certification body might need similar solutions, given expected size/number of stakeholders

AWS Sagemaker one option for streamlined handling of ML dev+testing (docker, environments, clusters). Gitlab has open-source, on-prem CI/CD license

Questions:

Do we need clinical/prospective evaluation IN ADDITION to the retrospective benchmarking?

Summary slides

https://docs.google.com/presentation/d/1uF8JEmQV\_4bjDVBlJiygrR\_fXTa4uTZyt8W79UQIfo8/edit#slide=id.p

OLD

https://drive.google.com/file/d/17GqJsFIhXJ4HJsZcuKOpTUpOpHoNa1hB/view?usp=sharing

Henry Notes on the Questions:

What are the requirements for an assessment platform?

dynamic TG specific metrics

informative, robustness, medical-performance

possibly case synthesisers for robustness cases

dynamic TG specific (hierarchical) of case and AI dimensions for reporting

e.g. case: age, sex, ethics, preconditions, region/location, skin type, social background, assessment-author, data-source, collection-date

e.g. for AI: offline/online, disease-subset, symptom-subset, number-of-presenting-complaints,

aware of the testing data statistics and its deviation from the stakeholder use case statistics

possibly with user-interface for checking it

base lines performance visualisation/comparison in the reporting

interactive result data drill-down / browsing

result publication/commenting/feedback management/UI e.g. for complaining about a certain case

(testing on stakeholder datasets)

registration/management of AIs incl. UI

an AI analysis page with its bias on all dimensions, errors, performances etc.

protect data and AI IP

protected against platform, AI and data manipulation

protected against malicious AIs

dialog-mode support

AI technology agnostic

result data storage/management

open source for transparency

complete GDPR compliance inc. tracing of consent and means for deleting cases if required etc.

How could a minimal working example for such a platform be implemented?

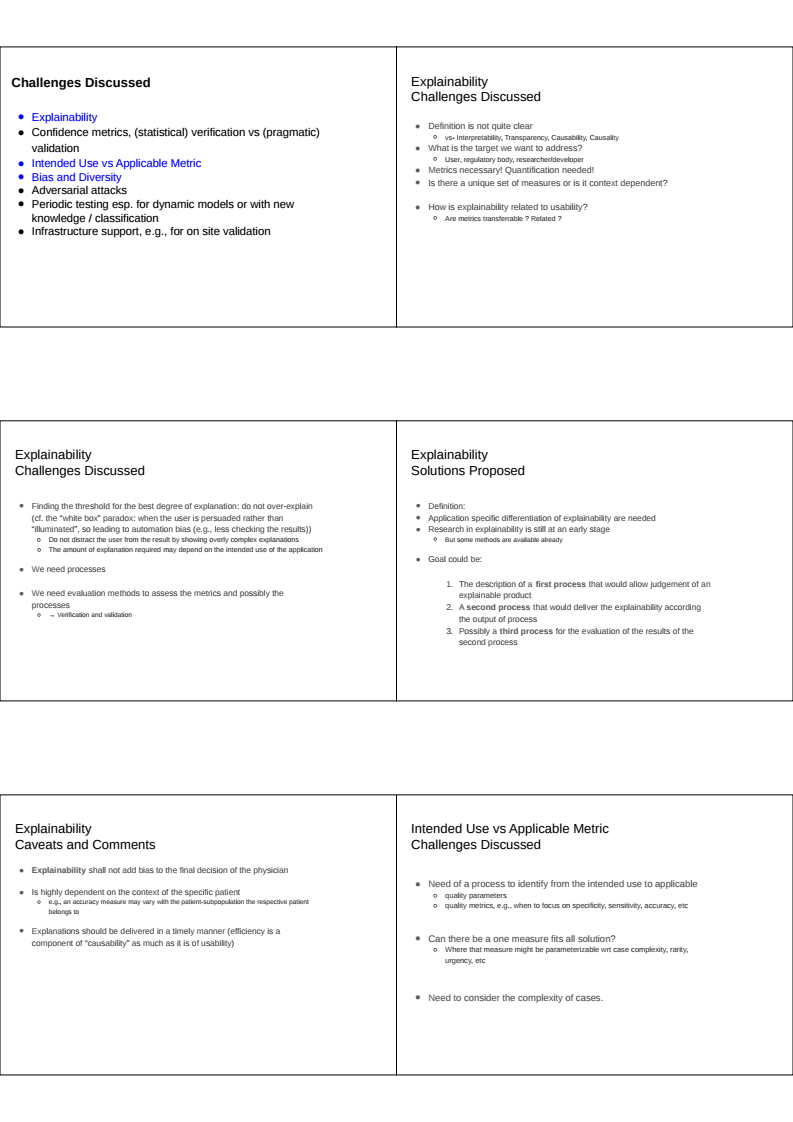
try to setup a challenge in AICrowd if possible e.g. using TG Snake as a template

create your own system if needed - maybe using TG symptom as pattern

maybe using atlas, pipelines etc.

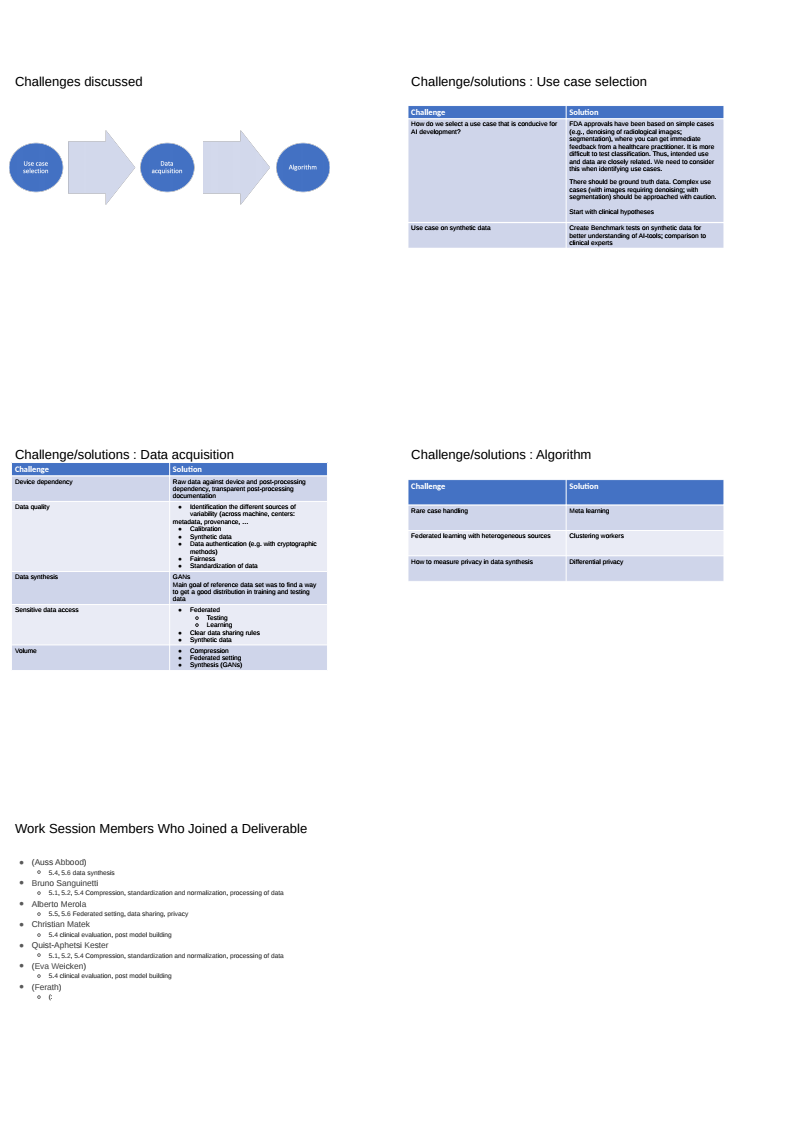
# Result Presentations from the Work Sessions Chairs

## Tests WS

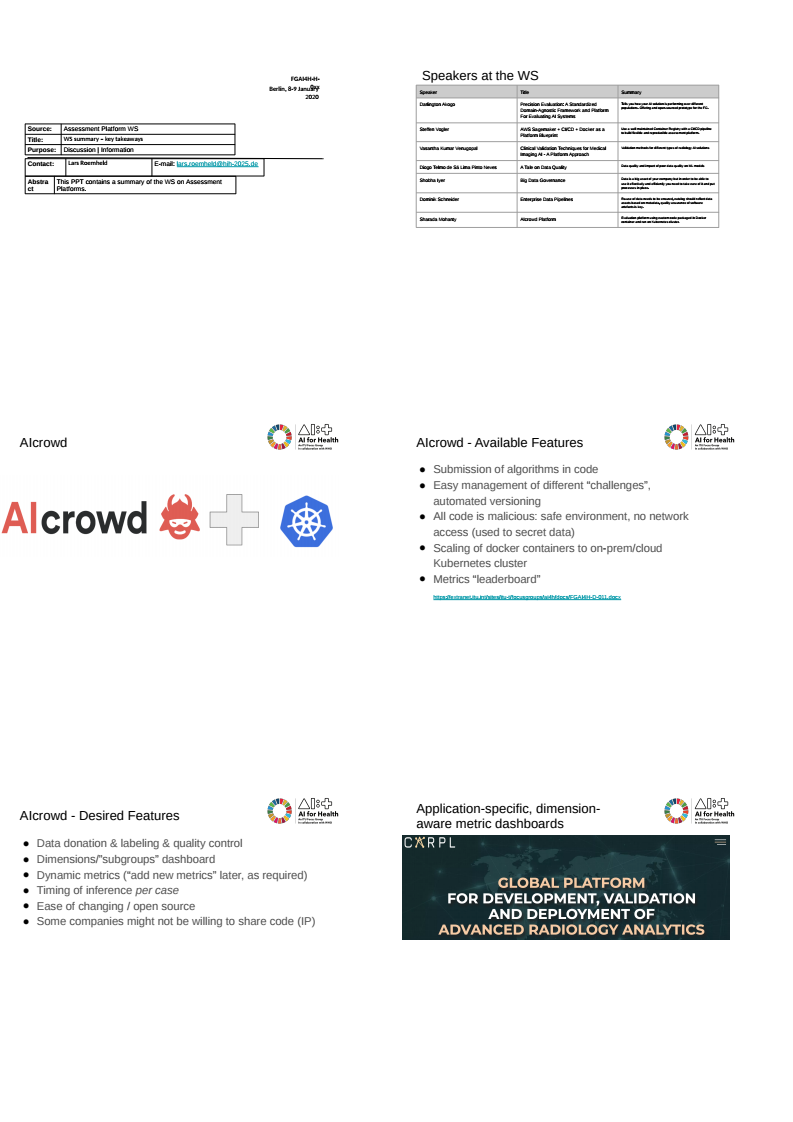
Original file: [*https://docs.google.com/presentation/d/1dtOk4NCiWki0xRJM64iMyt-aTIZbEPF3RTn-K0Vfi44/edit?usp=sharing*](https://docs.google.com/presentation/d/1dtOk4NCiWki0xRJM64iMyt-aTIZbEPF3RTn-K0Vfi44/edit?usp=sharing)

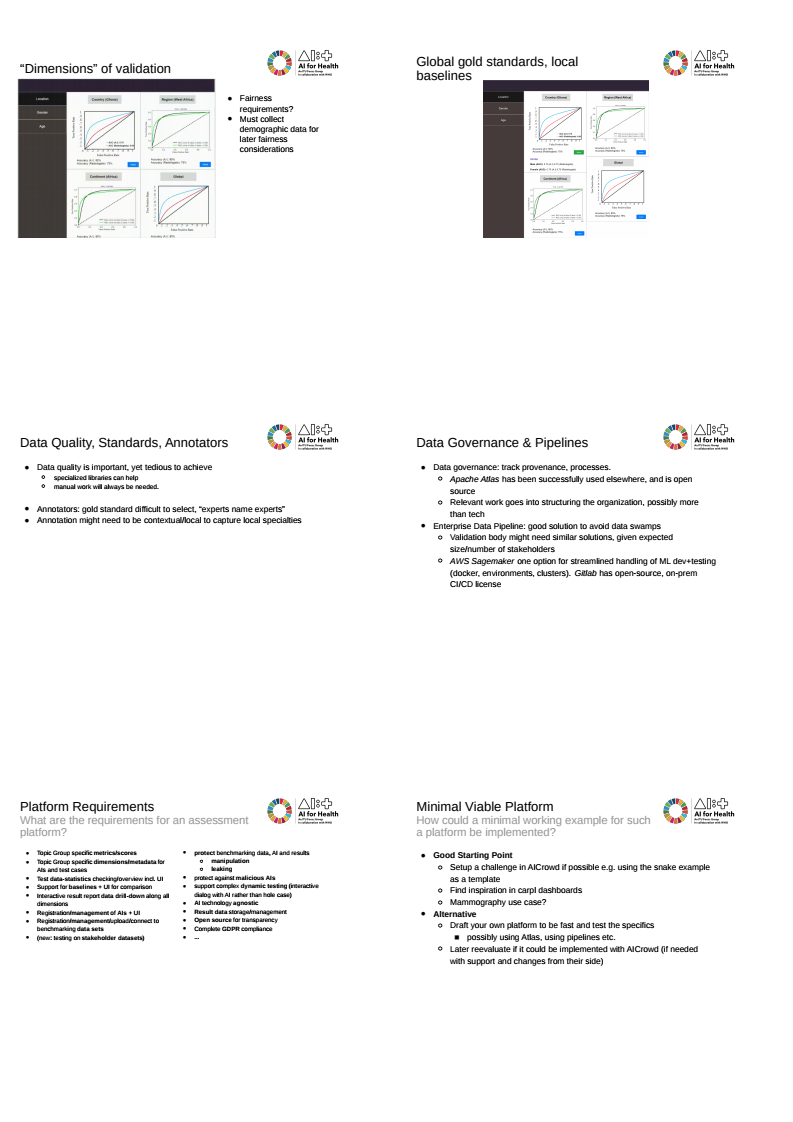
# 

## Data WS

Original file: [*https://docs.google.com/presentation/d/1-TfVH0H-FX29ES7mtYYAajd29MfBFHl39zlR\_8YddHo/edit?usp=sharing*](https://docs.google.com/presentation/d/1-TfVH0H-FX29ES7mtYYAajd29MfBFHl39zlR_8YddHo/edit?usp=sharing)

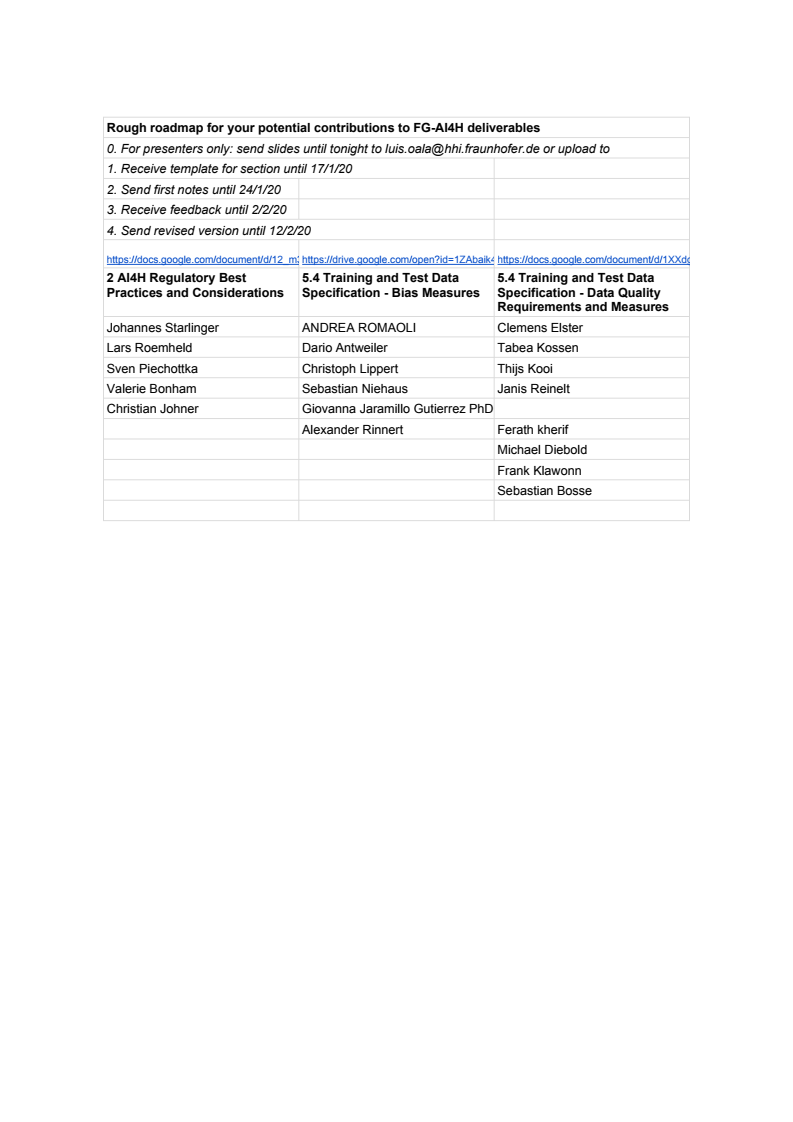
## Assessment Platform WS

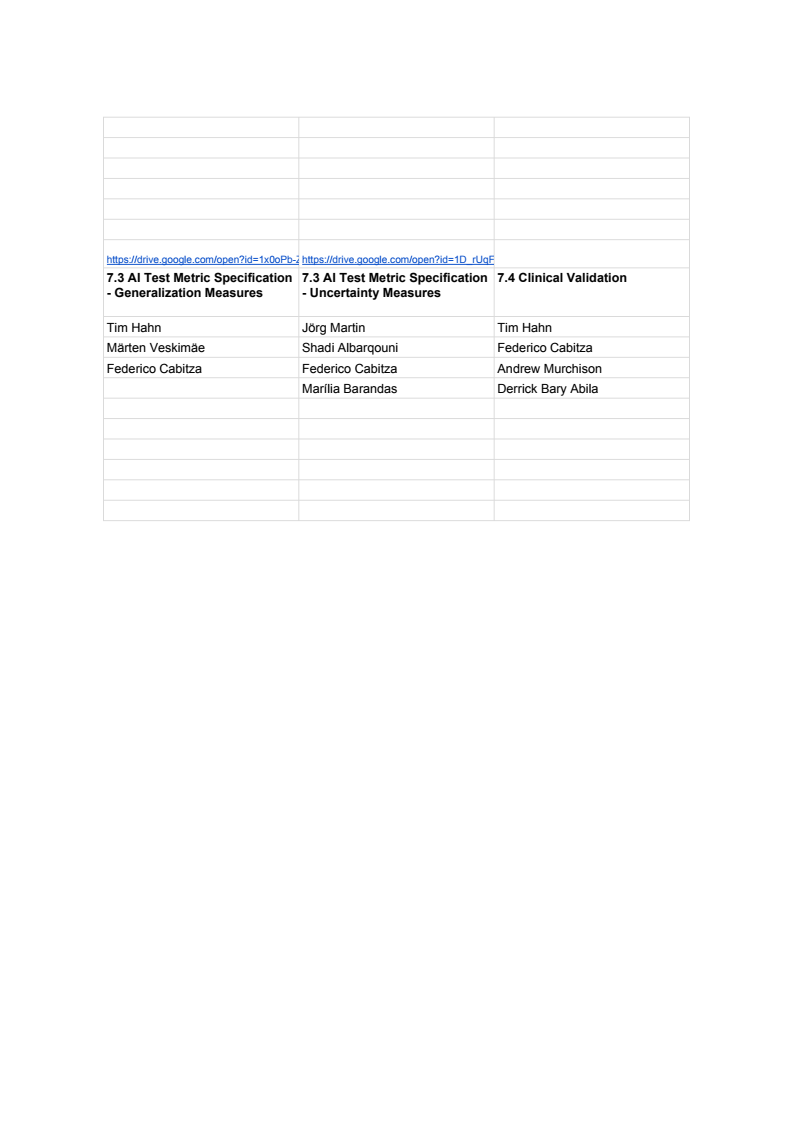
Original file: [***https://docs.google.com/presentation/d/1uF8JEmQV\_4bjDVBlJiygrR\_fXTa4uTZyt8W79UQIfo8/edit#slide=id.p***](https://docs.google.com/presentation/d/1uF8JEmQV_4bjDVBlJiygrR_fXTa4uTZyt8W79UQIfo8/edit#slide=id.p)



# Workshop Participant Volunteers for Contribution to FG-AI4H Deliverables

Original file: https://docs.google.com/spreadsheets/d/1u3p5QrqkArL8\_tJ8I1O5\_j3qYIeycYLP0TD0siWmfM4/edit?usp=sharing





\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_