Share Analysis. Not Data.

PRANA-DATA

DA.2 Proof of principle of an approach based on bringing the algorithms to the data - research setting

Project	PRANA-DATA
Project leader	Wessel Kraaij (TNO)
Work package	
Deliverable number	DA.2
Authors	Andre Dekker (MUMC)
Reviewers	Wessel Kraaij (TNO), Jessica Doorn (TNO)
Date	March 24, 2017
Version	2
Access Rights	Public
Status	Final

Summary

This document describes a proof of principle implementation in hospitals of a distributed learning infrastructure where research algorithms go to the data rather than the data to the research algorithm. Such an approach is considered to be privacy-by-design as without data leaving the hospital, privacy cannot be harmed. However, the approach also has downsides as an extreme level of semantic interoperability (so-called FAIR data stations) is required and not all research questions can be answered using this approach.

Contents

1 Introduction

The amount of data in hospitals in increasing rapidly (Figure 1[1]). This increase in data, often referred to as Big Data in health care is often seen as a positive thing as it might inform care professionals in personalize medicine and enable other aspects of P4 medicine[2].

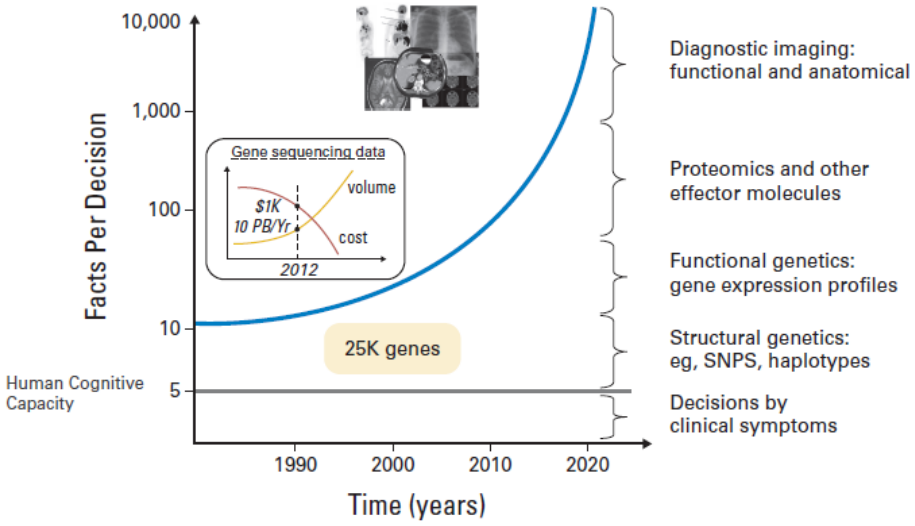
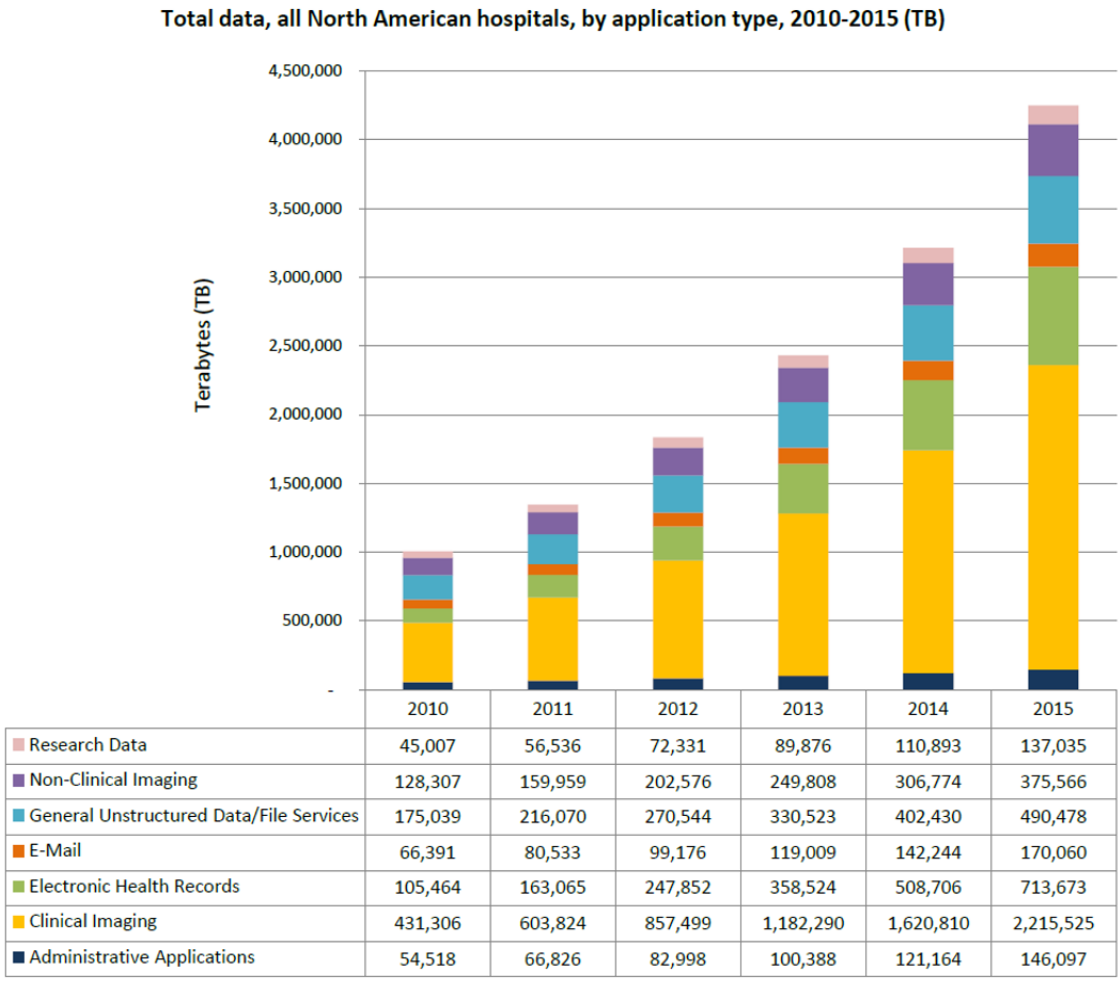


Figure 1: Rise in data elements in health care.

To give an estimate of the data which is available, following numbers for cancer are illustrative. In the last decade about 140 million persons worldwide are diagnosed with cancer. Suppose the data available on these patients range from 100MB in less economically develop countries up to 10GB in more economically develop countries such as the Netherlands. The total volume of data on cancer patients from the last decade is then between 14 and 1400 petabyte of data.

Another perspective is to look at North American hospitals as a whole. The estimate in 2011 was that in 2015 these hospitals hold about 4000 petabytes of data (Figure 2[3]). It is interesting to note that the majority of hospital data consist of clinical imaging data.



	2010	2011	2012	2013	2014	2015
Research Data	45,007	56,536	72,331	89,876	110,893	137,035
Non-Clinical Imaging	128,307	159,959	202,576	249,808	306,774	375,566
General Unstructured Data/File Services	175,039	216,070	270,544	330,523	402,430	490,478
E-Mail	66,391	80,533	99,176	119,009	142,244	170,060
Electronic Health Records	105,464	163,065	247,852	358,524	508,706	713,673
Clinical Imaging	431,306	603,824	857,499	1,182,290	1,620,810	2,215,525
Administrative Applications	54,518	66,826	82,998	100,388	121,164	146,097

Figure 2: North American hospital data.

However all this data far surpasses human cognitive capacity and thus new methods, e.g. artificial intelligence, data mining and decision support systems, are needed to process data and extract from the processed data meaningful and actionable information. Such a learning health care system in which evidence based medicine is complemented by data driven medicine is one of the ultimate goals of clinical data science research (Figure 3[1]). In short, data driven medicine tries to learn predictive models from routine clinical data and applies these models at the point of care in decision support systems.

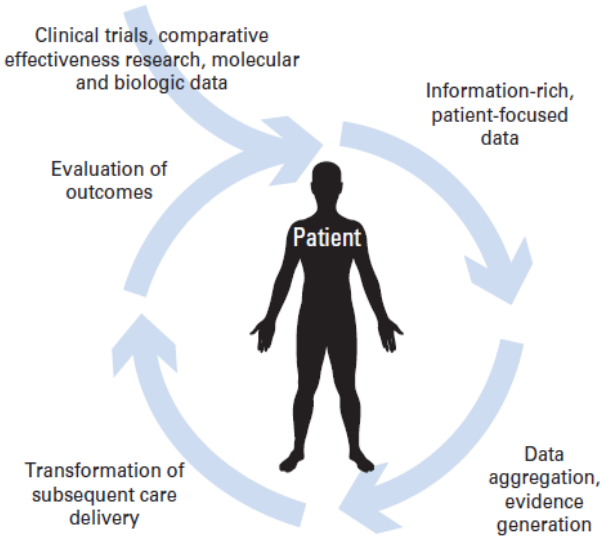


Figure 3: Rapid learning health care

For data driven medicine top work, access to hospital data is needed. However there are about 100.000 hospitals in the world, and the main challenge for data driven medicine is to get hospitals to share their data. Four main barriers to sharing data have been identified[4]:

1. Administrative (I don't have the resources)
2. Political (I don't want to)
3. Ethical (I am not allowed to)
4. Technical (I can't)

To solve these barriers, the Personal Health Train concept (PHT) was developed in which a radically different approach is taken. Unlike classical approaches (e.g. clinical trials, registries) data is not moved from the hospital to a central location where the research takes place. Rather the research algorithm is brought to the data answering the research questions locally, in each hospital, without moving the data, a concept called distributed learning.

Distributed learning addresses many of the barriers as hospitals no longer have to put resources (e.g. data managers) into filling in their data at a central location. They continue to have full control of their data to address the political barrier. As data does not leave the hospital, it is a privacy-by-design approach addressing the ethical barrier. Finally, the technical challenge of moving data around is no longer relevant as no data is moved.

We hypothesize that bringing algorithms to the data can answer research questions for data driven medicine. Specifically we think that distributed machine learning of predictive models for cancer care can be learned from hospital data without the need for data to leave the individual hospital.

The PRANA-DATA project aimed to investigate how privacy respecting analysis can be done on sensitive data that is distributed and should not be disclosed to the parties that perform the analysis. As part of the PRANA-DATA project, a proof-of-concept for distributed learning was implemented to learn a lung cancer outcome prediction model with the following aims

a) evaluate the proposed technology in a broader and more systematic setting and compare it to alternative privacy preserving methods;
b) describe the architecture, define value and identify shortcomings in the proposed technology;
c) evaluate the proposed technology in terms of abiding to FAIR principles, data protection regulations and process efficiency

2 Methods

The implementation of the PHT is based on the tools and ideas developed in the STW duCAT project and its associated projects (such as euroCAT, ozCAT, SAGE, chinaCAT).

Referring to green and red part in Figure 4 the system consists of two parts

1. Data preparation: The platform to extract data from clinical data sources and make them available for learning.
2. Distributed learning: The platform to do the actual learning of a predictive model.

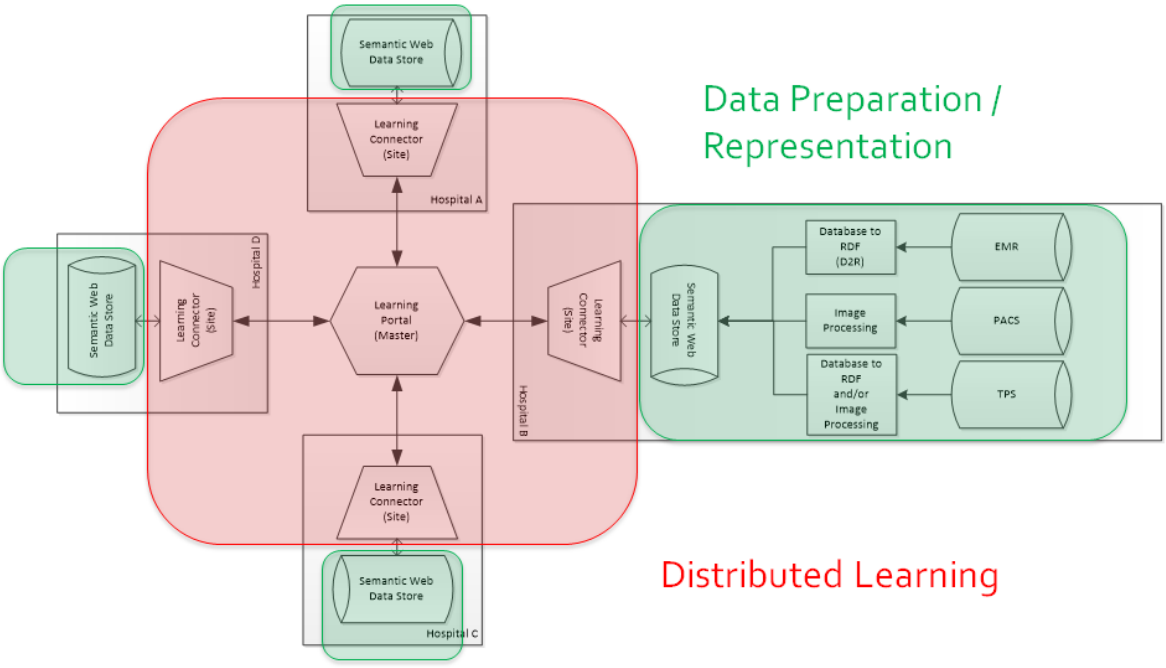


Figure 4: PHT implementation in two parts. A data preparation part and a data representation part.

2.1 Data preparation

Extracting data from clinical sources and make them available for learning is done using a set of free and/or open source tools. Referring to Figure 5, these tools first create de-identify data and make them syntactic interoperable. Syntactic interoperability is the level in which the data is available and can be read, is vendor neutrally stored but cannot yet be understood, e.g. it may still be in a local language. The next step is creating semantic interoperability. In this crucial step, the local data is mapped to ontologically well defined terms (e.g. using SNOMED, LOINC etc) and put on a Semantic Web page. Because everyone is forced to use the same ontology, applications can query the data in

each hospital without adapting the query to local vendors, language, customs etc. Nowadays data available with such a level of semantic interoperability is called FAIR data.

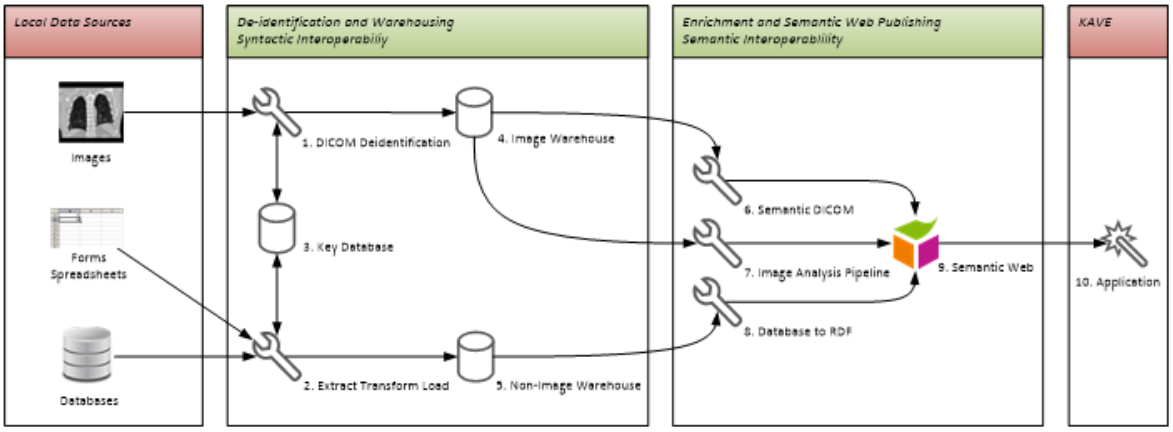


Figure 5: Converting local clinical data sources to linked data

Table 1: List of tools used in PHT proof of concept

Tool	Based on	Platforms	Functionality
1. DICOM Deidentification	CTP (RSNA)	Independent (Java)	Extract and de-identify DICOM objects according Supplement 142 of the DICOM standard. Load them into the Image Warehouse.
2. Extract Transform Load	Kettle (Pentaho)	Independent (Java)	Extract data from a multiple of text based sources, transform, de-identify remove and change data fields, load them into the Non-Image Data Warehouse
3. Key Database	PostgreSQL	Mult iple	Hold all sensitive information and identifiers for a patient. Generate random patient identifiers and provide these to DICOM Deidentification and ETL tools.
4. Image Warehouse	DCM4C HEE PostgreS QL	Independent(Java) Multiple	Receive, store and manage de-identified DICOM data in a file store and database. Transfer DICOM data to Semantic DICOM and Image Analysis Pipeline tools.
5. Non-Image Warehouse	PostgreSQL	Multiple	Receive, store and manage non-image data. Transfer non- image data to Database to RDF tools.
6. Semantic DICOM	SeDI	Independent (Java)	Convert DICOM header into Semantic Web publishable features (RDF). Store these features on the Semantic Web.
7. Image Analysis Pipeline	MIA	Independent (Java)	Analyse DICOM files and create Semantic Web publishable image derived features. Store these features on the Semantic Web.
8. Database to RDF	D2RQ	Independent (Java)	Query non-image warehouse. Create Semantic Web publishable features (RDF) by mapping local terms to the ontology. Store features on the Semantic Web.
9. Semantic Web	Blazegraph	Independent (Java)	Receive and store RDF from various tools in multiple graphs. Provide a SPARQL endpoint for the application.

The proof-of-principle evaluation in PRANA-DATA used installation of the above system at centers in Maastricht, Aachen (DE), Hasselt (BE), Liege (BE), Eindhoven (total of n=278) and Nijmegen (n=150) lung cancer patients treated with radiotherapy. While the first five installations were completed by MAASTRO/MUMC+ staff, the installation at Nijmegen was done during the project by Radboudumc staff so that an external experience with the tools was obtained.

2.2 Distributed learning - technical

The distributed learning infrastructure was developed in close collaboration with Varian Medical Systems (Palo Alto, CA, USA) and the Amazon cloud. The system allows the upload of application (e.g. a machine learning application) and subsequent distribution of that application to the hospitals that participate in the learning exercise. Then the system allows the local execution of the algorithms and communication of algorithm results between the hospitals. The latter is required to get a globally optimal model. In the infrastructure, a lot of emphasis is on privacy and security such as user-role based authorization, full control of hospitals in accepting/denying access, research agreements dealing with IP, liabilities and responsibilities, audit trails, transparent communication, certificate signed applications, etc.

2.3 Distributed learning - mathematical

Beside the technical infrastructure, distributed learning also has mathematical consideration. The main challenge is: How do you learn a globally optimal prediction model from data without putting all the data into one place? A number of solutions are available which can be categorized in parallel and sequential learning.

In sequential learning a model is learned in one hospital. The model is then sent to the next hospital which modifies the model according to their local dataset. The model is then sent to the third hospital, etc. Such sequential learning is especially suited for Bayesian approaches (see insert[5]), such as Bayesian Networks. The sequential learning approach was tested in the proof-of-principle implementation[6].

In parallel learning each hospital learns their own model, shares them and then a central "master" checks for convergence of the models and, if no convergence is yet reached, directs the hospitals to relearn a model with adjusted parameters.

Bayes's theorem

Beginning with a provisional hypothesis about the world, we assign to it an initial probability called the prior probability or simply the prior. After actively collecting or happening upon some potentially relevant evidence, we use Bayes's theorem to recalculate the probability of the hypothesis in light of the new evidence. This revised probability is called the posterior probability or simply the posterior. Specifically Bayes's theorem states that the posterior probability of a hypothesis is equal to the product of (a) the prior probability of the hypothesis and (b) the conditional probability of the evidence given the hypothesis, divided by (c) the probability of the new evidence.

From: Paulos JA: The Theory That Would Not Die - By Sharon Bertsch McGrayne - Book Review. The New York Times , 2011

There are a number of known solutions for parallel learning. One method, implemented in the proof-of-principle, is the alternating direction method of multipliers (ADMM)[7, 8]. With ADMM a number of popular statistical and machine learning models can be learned including lasso, linear and

logistic regression and support vector machines. Additionally we tested distributed learning of random forest models, another popular machine learning model.

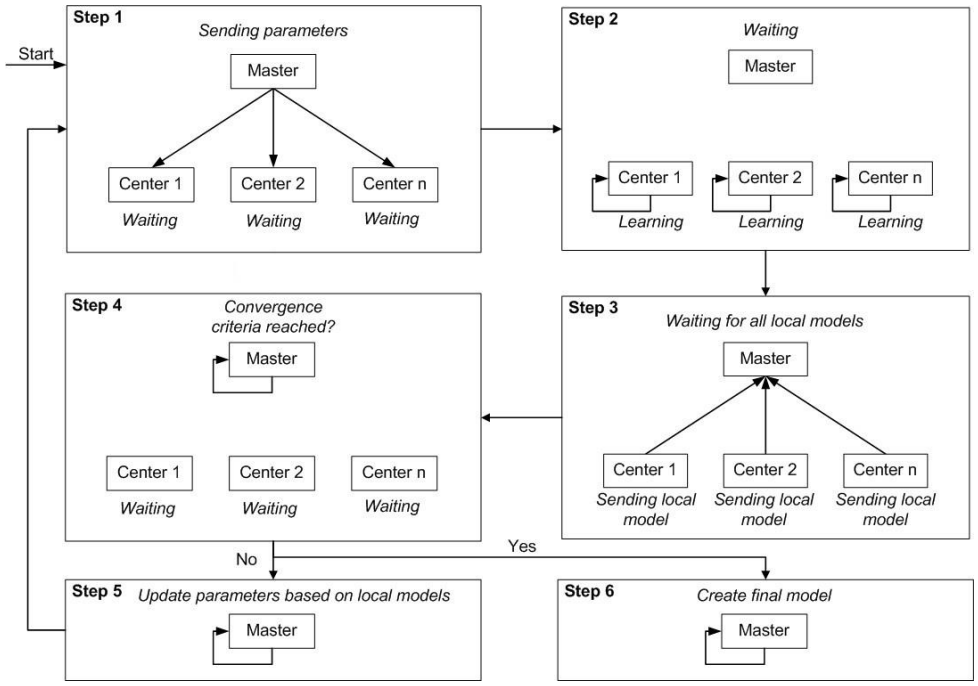


Figure 6: Parallel distributed learning

3 Results

Two models were learned using distributed learning, both predicting dyspnea (shortness of breath) after radiotherapy for lung cancer. One was a Bayesian Network the other a Support Vector Machine.

3.1 Bayesian network

The Bayesian Network was learned in sequential distributed learning. This network is given in Figure 7. The AUC (a measure of performance) of this network was 0.61 which is moderate but compared well to previous models.

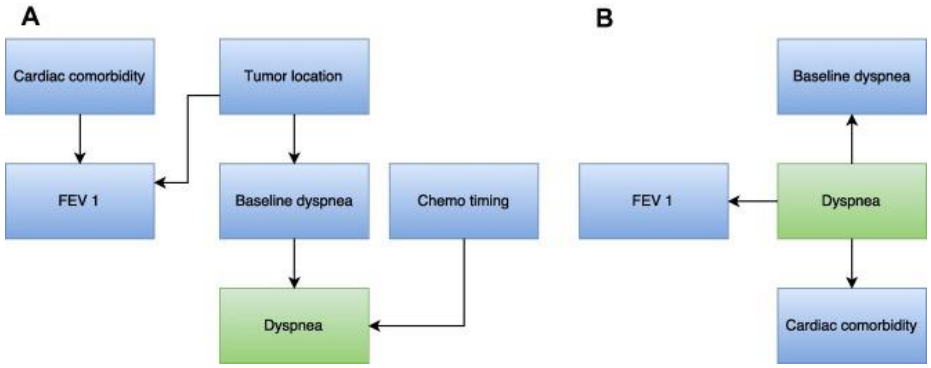


Figure 7: Bayesian network for dyspnea

To answer the question if a model learnt in distributed learning has a worse performance than a model learnt in the classical way (by centralizing the data), we performed a comparison analysis with bootstrapped samples of 100, 1.000 and 10.000 patients (Figure 8) with different levels of degradation (by introducing missing data) and with different numbers of hospitals.

The results show that difference in the Bayesian network learned and in its performance are minimal for large numbers of patients, with high quality data distributed across a small number of hospitals. Even for quite degraded data spread around many hospitals the decrease in performance was acceptable.

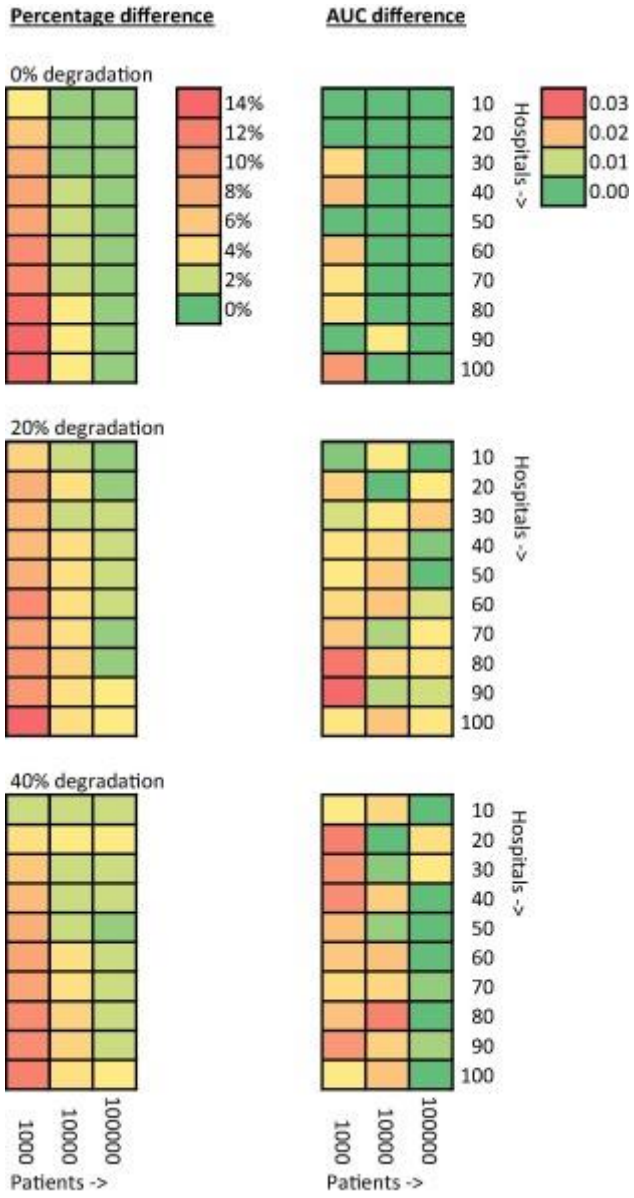


Figure 8: Comparison of distributed learning versus local learning. Datasets were created by random sampling from the MAASTRO clinic data (N = 123). The first column shows the average difference in percentages of the conditional probability tables for the global and distributed model. The second column shows the difference in AUC for the global and distributed model. Rows depict the levels of artificially introduced random missing data (0%, 20%, 40%, respectively).

3.2 Support Vector Machine

The support vector machine model was learned in parallel distributed learning using the ADMM method described above. Again the outcome predicted was dyspnea after radiotherapy. The model had an AUC of 0.66, again comparable to previous results.

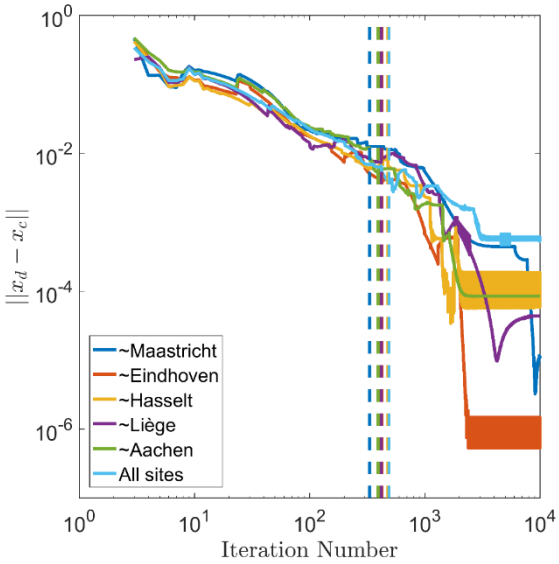


Figure 9: Convergence graphs of distributed ADMM solutions x_d to centralized solutions x_c for 10^4 iterations. Vertical lines indicate the iterations in which internal convergence criteria were met in the distributed learning network. The data was created in local simulations. ~ indicates 'Trained on all sites except'.

Comparing to a centralized solution, it can be seen that the model converges in all cases to the central model, something which can also be proven mathematically[8]. The parallel solution with 500 iterations and five nodes took about 2 hours to run in 5 centers.

4 Conclusions

The main conclusion from this work is that privacy respecting analysis can be done on sensitive data that is distributed without disclosing the data to the parties that perform the analysis. We implemented this distributed learning as a proof-of-concept and met the aims set out beforehand.

4.1 Comparison with other privacy preserving approaches

Specifically we evaluated the proposed technology in a broader and more systematic setting than ever before. The main comparison in PRANA DATA is between distributed learning and homomorphic encryption. In general, we can conclude that distributed learning is a somewhat easier technology to implement than homomorphic encryption. It does not require a large effort from the researcher like changing the machine learning library to 'handle' encrypted data. Rather, standard libraries in Matlab, R, Java etc. can be used for distributed learning. On the other hand, the mathematics of distributed learning are somewhat more cumbersome as beside the local learning one needs to consider merging the locally learned model e.g. using ADMM or Bayesian approaches.

4.2 Description of architecture, value and limitations

Another specific aim was to describe the architecture, define value and identify shortcomings in the proposed technology. This document describes the architecture at a high level, with details available in publications and in press manuscripts. The value of such an architecture is its relative simplicity and clarity in terms of roles and responsibilities. The drawbacks is that this approach is likely limited to certain classes of machine learning. Although not all possible algorithms have been evaluated, it is likely learning algorithms with non-convex optimization characteristics such as neural networks and the associated deep learning are less suitable for distributed learning and more for a homomorphic encryption approach.

Also, we exclusively looked at the problem of horizontally partitioned data, meaning data distributed in such a way that two holders have separate but individually complete data set (e.g. hospitals having each their own patient population). A problem not tested is that of vertically partitioned data, where one holder has some data elements on a patient population and another holder has additional data on the same population (e.g. a hospital having the treatment data and a general practitioner having the follow-up data). Learning across vertically partitioned data is still an area of active research.

Finally, the architecture requires centers to make their data available as FAIR data. The current tools still require significant investment (3 person months for a limited set of data element for lung cancer patients). We feel that improvement in these would open the way for more centers to join and thus more data to become available for learning.

4.3 FAIR principles, data protection and efficiency

A final aim of the PRANA DATA proof-of-principle was to evaluate the proposed technology in terms of abiding to FAIR principles, data protection regulations and process efficiency.

The proposed methodology and tools to make hospital data available for learning is considered to be FAIR avant-le-mot. The linked data approach makes sure data is interoperable and reusable. Accessibility and findability are guaranteed using the distributed learning infrastructure. So is our approach really FAIR? This depends on the exact FAIR conditions which are still in development, but we can say that it is certainly close to FAIR data.

The approach taken in the distributed learning proof-of-principle is a principled privacy-by-design approach. Current data protection and research regulations have not considered this approach, as evident by data protection and research review boards having difficulty ruling on the approach as it seems to be outside their remit. It remains to be seen if the proposed approach is truly accepted as falling within the new GDPR regulation proposed by the EU and implemented in the Netherlands in 2018. An opinion by the AP (Autoriteit Persoonsgegevens) in this matter will be sought in due time.

With regard to process efficiency, in the current implementation learning from 5 centers takes about 2 hours in parallel learning with the center with the most limited hardware determining the pace. With engineering dedicated at improving efficiency, it is expected that the time to learn a model can be reduced significantly, making distributed learning a realistic approach in terms of process efficiency. The reported comparison between a model learned in a distributed fashion and a model

learned centrally show that distributed learning can equal the quality and performance of models learned centrally.

5 References

1. Abernethy AP, Etheredge LM, Ganz PA, et al: Rapid-learning system for cancer care. J Clin Oncol 28:4268–4274, 2010

2. Hood L, Friend SH: Predictive, personalized, preventive, participatory (P4) cancer medicine. Nat Rev Clin Oncol 8:184–187, 2011

3. John McKnight, Brian Babineau: North American Health Care Provider Information Market Size & Forecast. Enterprise Strategy Group, 2011

4. Sullivan R, Peppercorn J, Sikora K, et al: Delivering affordable cancer care in high-income countries. Lancet Oncol 12:933–980, 2011

5. Paulos JA: The Theory That Would Not Die - By Sharon Bertsch McGrayne - Book Review [Internet]. N Y Times , 2011[cited 2017 Mar 21] Available from: http://www.nytimes.com/2011/08/07/books/review/the-theory-that-would-not-die-by-sharon-bertsch-mcgrayne-book-review.html

6. Jochems A, Deist TM, Soest J van, et al: Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiother Oncol 121:459–467, 2016

7. Boyd CA, Benarroch-Gampel J, Sheffield KM, et al: 415 patients with adenosquamous carcinoma of the pancreas: a population-based analysis of prognosis and survival. J Surg Res 174:12–19, 2012

8. Damiani A, Vallati M, Gatta R, et al: Distributed Learning to Protect Privacy in Multi-centric Clinical Studies [Internet], in The 15th Conference on Artificial Intelligence in Medicine. Pavia, Italy, Springer, 2015[cited 2015 Jun 26] Available from: http://eprints.hud.ac.uk/23905/