Pilot 5: Heart Failure

1. Key Information

1.1 Involved Partners

- Erasmus Universitair Medisch Centrum Rotterdam (EMC)
- Achmea BV (ACH)
- Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek TNO (TNO)

1.2 Involved Countries

- Netherlands

1.3 Keywords

- Multi-Party Computation
- Encryption
- Secure lasso regression
- Prediction model
- Heart Failure

1.4 Task Description

The goal of this pilot is to identify comorbidities that have the strongest correlation with the number of hospitalizations in context of heart failure. These comorbidities are then used to guide an intervention that is designed to reduce the number of hospitalizations. A Multi Party Computation (MPC) system is developed that allows for the secure combined analysis of datasets residing in the hospital databases and health insurance databases. The key feature of this approach is that it allows a secure analysis of two datasets that the owners cannot share with the others.

1. The data is collected and is loaded into the MPC application that is installed at a laptop, server or VM at both Achmea and Erasmus MC.
2. The MPC tool encrypts all data. Each party uses its own public key for encryption and keeps the private key secret.
3. The MPC tool of the different parties (Achmea, Erasmus MC and ZorgTTP) exchange encrypted data in such a way that the prediction model can be computed without revealing anything else about their data.
4. One or more of the parties receive the computed prediction model.
5. Before, during or after the computation process, no input data of one of the parties is received by any other party (only in encrypted form, which cannot be decrypted by the receiving party).

2. Building Blocks

2.1 Architecture

2.1.1 System Architecture

Figure 1 displays the existing architecture at Achmea that is used for the analysis:

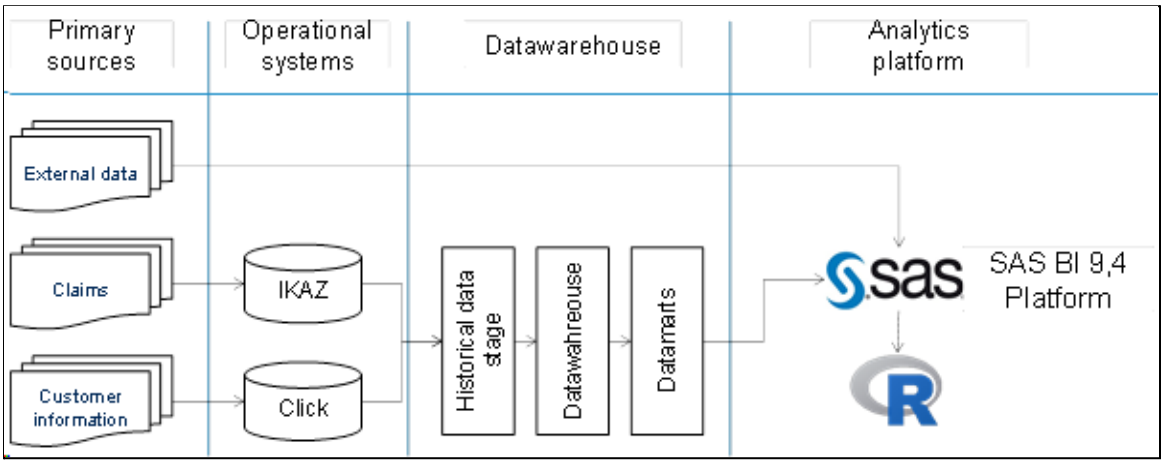

Figure 1

In the Netherlands most (99%) healthcare procedures are claimed digitally by the care provider to the healthcare insurers. An additional stream of data with personal data (name, data of birth, living address, etc.) are received from the “Basisregistratie Personen (BRP)” (national administration about all Dutch citizens managed by the government).

All the data is collected at ACH in a separate operational system. The data is made available to the organization with a data warehouse (Oracle) to different types of applications. The quality standard of the data warehouse is very high, because it is also used for mandatory reports to the government. For this project the SAS BI Platform and R can be used to extract the data and analyse the data. The analytical platform makes it also possible to integrate external data into the analysis.

2.1.1.2 Multi Party Computation

Secure learning with secure MPC

The secure learning solution consists of two parts: the hidden set intersection and the regression analysis. The hidden set intersection method finds matching persons in the database of Erasmus MC and Achmea in a secure way. With this matching information a secure regression model can be trained on the combined features, again while not disclosing any

information on the individual persons. Both algorithmic steps are based on secure multi-party computation.

Hidden set intersection

Firstly, the two data providers (Erasmus MC and Achmea) need to determine which patients are present in both their datasets, the so-called set intersection. Computing the intersection with MPC ensures that no party knows which patients are in the intersection. Only the size of the intersection will be learned by both parties, but not which (individual) patients. This is called hidden set intersection. The state-of-the-art MPC protocol involves a third party (e.g. ZorgTTP) that helps the parties to efficiently determine this. This third party also does not learn anything but the size of the intersection. The (encrypted) data is stored locally at each party and (encrypted) data is exchanged via https (please also figure 2).

Figure 2

Secure Linear Regression

Secondly, after securely retrieving the encrypted set of patient features in both datasets from Achmea and Erasmus MC, the secure regression analysis can be performed. At this stage, there is also a third party that helps in securely analysing the data. It is similar to the previous step (hidden set intersection) with the addition that there is also communication between Data Provider A and Data Provider B (Figure 2). Again, the (encrypted) data is stored locally at each party and (encrypted) privacy sensitive patient data is exchanged via https.

The secure regression analysis is implemented with an MPC protocol in Python and uses the open source library MPyC extensively, which is based on Shamir Secret Sharing. Communication in MPyC is performed by sending messages to specified IP addresses and port numbers of the involved parties; the library OpenSSL is used to guarantee the security of the communication. The hidden set intersection is a secret shared among the parties and because of the underlying mathematical properties of secret sharing, new values can be computed from the data. Because of the specific secure treatment of the data, the fitting of the regression model cannot be done using standard regression tools (such as in the Python toolbox scikit-learn). Instead, for the optimisation and fitting of the regression model we use Gradient Descent algorithms. In each iteration, the entire dataset (i.e. the obtained intersection) is used to compute a new gradient. However, depending on the method (stochastic, mini-batch or batch optimisation), in each iteration, the algorithm computes new gradient values on a subset of the data until the entire dataset is covered in sub iterations. The sub setting of the data is securely done using an MPC protocol and no party learns which subset of the data is used to compute the new gradient. As such, parties cannot learn new information about the data. Hence, no other information than the model coefficients is revealed to the parties.

2.1.2 Data Flow & Interoperability of services

2.1.3 Necessary Hardware

To run the MPC application, at each partner site the hardware requirements for the MPC installation are as follows: A Linux operating system (or another OS that can use Docker virtualization) is needed running on a system consisting of 60Gb disc, 4 gig memory, 4 cores and open ports: 220, 80 en 443. In additon, direct access or remote access with exportable graphics to a PC (in the local network) is required and the system should be externally reachable with reasonable bandwidth for significant communicational load with opened port (443) on the firewall to make the server externally reachable via HTTPS.

2.1.4 Software Components

The installation of the MPC software consists of the installation of three Docker containers: HTTPS client, HTTPS server and MPC module including graphical interface.

2.3 Data Processing

2.3.1 Processing of large structured / unstructured data sources

2.3.1.1 Data Sources

A. Comorbidity analysis of heart failure patient hospitalizations

How was data prepared:
Data is available from the Data Warehouse of Zilveren Kruis. These data are highly reliable because of national standards and because we have to justify our claims to the government. Data cleaning was therefore no issue. We took into account the implications of data artefacts because of system changes.

How was data integrated:
Claims data and patient characteristics are available on an individual level. External data was joined on the level of postal codes.

How was data analysed:
Based on LASSO regression we predicted hospital admission out of > 4.000 features, based on an individual level.

B. Multi Party Computation

Two datasets are used to test the models computed with the MPC solution. Generated data, based on the datasets description by EMC and ACH, and an open data set about diabetes patients. In case data from the data providers ACH and EMC is used, the data cleaning process will need to take place locally at each party before running the protocol.
The secure regression analysis contains three models: Linear regression and Ridge regression. Each of these models are trained with the use of Gradient Descent algorithms, either stochastic, batch or mini-batch gradient descent. The Gradient Descent algorithms can be seen as the building blocks behind the regression analysis, solving the optimization problem. However, not all algorithms can work with Gradient Descent method. The goal is to work towards LASSO regression, which can be built on the Ridge regression. This LASSO regression method aims at finding sparse regression models that enable easier model explanation, because fewer non-zero coefficients will be obtained. The output of the MPC solution is the coefficients of the regression model, such that the parties learn the relation between the different features.

Data Source	**Description**	**Acquisition**	**Characteristic (Size, Patients, Years, Origin/Region)**

Multiple sources	**Integration to data warehouse**	**Data access**	**Data stored in cloud**	**Multi-party architecture**	**Secure environment**	**Transform raw / unstructured data**
yes	unsure	access of de-identified data through firewall of institution using secure login	no	yes	yes	yes

2.3.1.2 De-Identification and anonymisation

2.3.1.3 Acquisition

We will use readily available data in existing databases at Achmea and Erasmus MC.

Due to confidentiality, the MPC will work with generated datasets from both Achmea and EMC. In parallel to the proof-of-concept demonstration, the next steps towards operationalization of the MPC analysis will be investigated.

2.3.1.4 Cleansing

The patient data will be cleaned with regards to our inclusion criteria for chronic HF. We will make sure the data is complete and normalized per feature.

For the MPC setting, the data cleaning will happen locally at each party (Achmea and EMC).

The MPC solution will deal with data integration and apply a regression analysis on the patients that are available in both datasets. This will be done in another MPC solution, so that it remains hidden which patients are in the intersection of both datasets.

2.3.1.5 Data Integration

The data will be integrated on the patient level. We are making use of MPC to allow for secure data integration and analysis. This approach allows us to securely and in an anonymized fashion combine and integrate databases from different owners who do not wish to share their databases with each other.

The MPC solution makes sure that the data input of both EMC as Achmea remains private. A trusted third party (TTP) is added to the MPC scheme in order to speed up the MPC computations. Note that this TTP does not learn anything about the private data inputs of the other parties, but only assists in the computations. The data integration will be performed by a hidden set intersection, where no one learns the identity of the shared data records (patients) and yet a regression analysis can be performed on the intersection of the data. The outcome of the MPC protocol are the weights for the LASSO regression coefficients, so that all involved parties learn the relation between various life style parameters and the number of hospitalizations.

2.3.1 Multi-velocity processing of heterogeneous data streams

Does not apply

2.3.5 Complex real-time event detection

Does not apply.

2.3.5.1 Notifications

2.3.5.2 Situations of Interest

2.3.5.3 Event Processing

2.3.5.4 Event Sources

2.3.5.5 Evaluation

2.4 AI Components

2.4.1 Deep learning for multilingual NLP and image analytics

2.4.1.1 Natural Language Processing

Does not apply

2.4.1.1.1 Evaluation

2.4.1.2 Image Processing

Does not apply.

2.4.1.2.1 Evaluation

2.4.2 Prediction Algorithms

2.4.2.1 Task

Prediction of HF patient hospitalizations

2.4.2.2 Data, Data Modelling

2.4.2.3 Features

1. Patient claims data
2. Patient characteristics

- EMR
    - Socio-demographic
    - Structured
        - lab, diagnoses, medications, hospital stay
- Insurance

2.4.2.4 Model

Within the HF pilot we are using databases containing a large number of HF patients and their many comorbidities. Our key KPIs is a reduction in the number of hospitalizations and we are planning on using machine learning approaches to identify those comorbidities that have the strongest correlation with the number of hospitalizations. Based on identified comorbidities we will design an intervention that will be tested in a prospective study on patients that fit our inclusion criteria. For details on the architecture for data storage and analysis see the below figure 3:

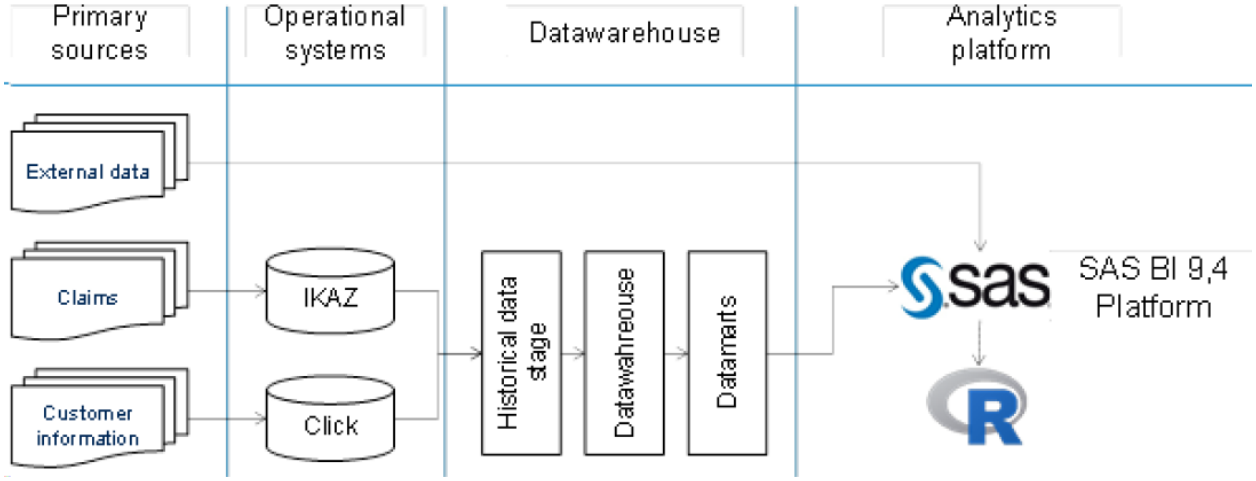

Figure 3: Details on the architecture for data storage and analysis

We are planning on using different types of machine learning solutions to investigate which ones are the most informative for our problem. Some of the machine learning approaches we will apply are:

• Random forest
• LASSO
• Elastic net.

2.4.2.5 Evaluation

2.5 Security and privacy of data access and processing

There are currently 15 million HF patients in Europe. The annual risk of death for these patients is 10% to 20%. In particular patients with multiple comorbidities do not currently receive the optimum medical treatment, which leads to many hospital admissions. Achmea, Erasmus MC and TNO work together in this pilot to offer personalized medical treatments based on secure data analyzes. Multi-party computation (MPC) is used for this, where analyzes are carried out on the combination of data from Achmea and Erasmus MC without harming the privacy of individuals.

The MPC solution consists of two parts:

1. Hidden set intersection. The Achmea and Erasmus MC Rotterdam Study datasets both contain of a group of patients and attribute values are known to either Achmea or Ergo. For the analysis following this step, it is important to use the overlap of patients in the two groups. That is, the group of patients that appear in both the Achmea and Ergo data. This is also called the intersection of the two datasets. The intersection itself remains concealed during the hidden set intersection; only the encrypted set intersection is computed. The only thing that becomes known during the hidden set intersection is the size of the intersection, so the number of patients that occurs in both datasets (in the intersection). The result of the hidden set intersection is an encrypted dataset of the patients that appear in the data from both Achmea and Ergo. Encrypted means that nothing is known about these patients or their data.
2. Secure LASSO regression. The encrypted data of patients in the intersection set is used to find a statistical relationship between influencing factors (such as: number of cigarettes per day, weight, medication compliance, etc.) and the number of hospitalization days. During the secure LASSO regression, no information is revealed other than the outcome of the regression. The outcome of the secure LASSO regression is the regression coefficients that indicate the linear relationship between the influence factors and the number of hospitalization days.

The results of the secure LASSO regression can be used to advise (new) HF patients on the influence of their behaviour (smoking, exercising, etc) on the expected number of hospitalization days. Instead of the current qualitative advice, doctors can give quantitative advice on lifestyle and medicine compliance.

2.5.1 Access Control

2.5.1.1 Authentication

Not applicable for the MPC proof of concept.

2.5.1.2 Authorization

The MPC computation uses TLS/SSL protocols for communication. The SSL certificates are generated and distributed to all parties beforehand. Each party can check the certificates of the other parties to authenticate. Additionally, only specific IP addresses are allowed to join the MPC protocol.

2.5.2 Data Protection

2.5.2.1 Data at rest

During the MPC computation, the involved data will remain private to each partner. The data and outcome of computations that are performed are encrypted by means of Shamir secret sharing and Paillier additive homomorphic encryption. Homomorphic encryption. An (additively) homomorphic encryption scheme is a public-key encryption scheme, thus consisting of a key-generation algorithm KeyGen, an encryption algorithm Enc and a decryption algorithm Dec. For a keypair of public and secret key (pk; sk) generated by KeyGen, we have that Encpk takes as input a message m and some randomness r, and produces as output a ciphertext c = Encpk(m; r), with the property that Decsk(c) = m, and that no information whatsoever can be extracted on m or sk from c and pk; in formal terms, the encryption scheme is IND-CCA1 secure. In order to simplify notation, we will often omit the key and randomness when discussing encryption, and write [m] := Encpk(m; r); moreover, we implicitly assume messages to be numeric values, so that addition and subtraction of messages are well-dened. The scheme is supposed to be additively-homomorphic, which means that there exists special operations on ciphertexts and , such that [m1][m2] = [m1 + m2], and [m1][m2] = [m1 m2] for all messages m1;m2. Our secure inner join protocol between AC, EMC, and ZorgTTP uses cryptographic hash functions, and both AC and EMC have an (additively) homomorphic encryption key pair; we used SHA-256 [37] as hash function and the Paillier homomorphic-encryption scheme [38] in our implementation. The (pseudo) randomness is implemented by PEP 506. That means that nothing can be derived from the (secret shared) data of one party.

2.5.2.2 Data in transit

The data in transit is encrypted as described above. In addition to that, the encrypted data exchanges are sent using TLS/SSL.

2.5.3 Auditory and logs

2.5.3.1 System Auditory

For the MPC proof of concept demonstration, there will be no application or services logs in place.

2.5.3.2 Services Auditory

For the MPC proof of concept demonstration, there will be no system logs in place.

2.5.4 Privacy measurements

2.5.4.1 Data Privacy Impact Assessment (DPIA)

2.5.4.2 Legal/Ethical process

2.5.4.3 Processes for complying with the current legislation

2.6 Trustworthy AI

2.6.1 technology/user adoption and establishing trust

2.6.2 ethical principles

- respect for human authority
- prevention of harm
- fairness
- explicability

2.6.3 key requirements

- Human agency and oversight
- Technical Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

2.7 System-Interaction

2.7.1 Human-Machine Interface / GUI

2.7.2 Education

3. Learnings

3.1 Challenges & Barriers

- Architecture
- Processing of large structured / unstructured data sources
- Multi-velocity processing of heterogeneous data streams
- Complex real-time event detection
- Natural Language Processing
- Image Processing
- Prediction Algorithms
- Security and privacy of data access and processing
- Trustworthy AI
- System-Interaction

3.2 Lessons Learned

- Architecture
- Processing of large structured / unstructured data sources
- Multi-velocity processing of heterogeneous data streams
- Complex real-time event detection

- Natural Language Processing
- Image Processing
- Prediction Algorithms
- Security and privacy of data access and processing
- Trustworthy AI
- System-Interaction

3.3 Main (quantifiable) achievements

4. Output

4.1 Papers

Impact of sex differences in co-morbidities and medication adherence on outcome in 25 776 heart failure patients.

Gürgöze MT, van der Galiën OP, Limpens MAM, Roest S, Hoekstra RC, IJpma AS, Brugts JJ, Manintveld OC, Boersma E. ESC Heart Fail. 2021 Feb;8(1):63-73. doi: 10.1002/ehf2.13113. Epub 2020 Nov 28. PMID: 33247631 Free PMC article.

Heart failure and promotion of physical activity before and after cardiac rehabilitation (HF-aPProACH): a study protocol.

Limpens MAM, Gürgöze MT, Lenzen MJ, Roest S, Voortman T, Kavousi M, Ter Hoeve N, Sunamura M, den Uijl I, van der Spek PJ, Brugts JJ, Manintveld OC, IJpma AS, Boersma E. ESC Heart Fail. 2021 Jul 16. doi: 10.1002/ehf2.13505. Online ahead of print. PMID: 34268900 Free article.

Veugen T, Kamphorst B, L'Isle N van de, Egmond MB van. Privacy-Preserving Coupling of Vertically-Partitioned Databases and Subsequent Training with Gradient Descent. Cyber Security Cryptography and Machine Learning. 2021; 38-51. doi: 10.1007/978-3-030-78086-9_3.

Egmond MB van, Spini G, Galien, Onno van der, IJpma A, Veugen T, Kraaij W, Sangers A, Rooijakkers T, Langenkamp P, Kamphorst B, L'Isle N van de, Kooij-Janic M. Privacy-preserving dataset combination and Lasso regression for healthcare predictions. BMC Medical Informatics and Decision Making. Accepted, to be published in September 2021.

Veeningen M, Chatterjea S, Horváth AZ, Spindler G, Boersma E, van der Spek P, van der Galiën O, Gutteling J, Kraaij W, Veugen T. Enabling Analytics on Sensitive Medical Data with Secure Multi-Party Computation. Stud Health Technol Inform. 2018;247:76-80. PMID: 29677926.

Prediction of long-term hospitalization and all-cause mortality in patients with Chronic Heart Failure on Dutch claims data: a machine learning approach. Onno P. van der Galiën[a], MSc; René C. Hoekstra[a,e], MSc; Muhammed T. Gürgöze[b], MD, MSc; Olivier C. Manintveld[b], MD, PhD; Mark R. van den Bunt[a], PhD; Cor J. Veenman[c,d], PhD; Eric Boersma[b], MSc, PhD, FESC , submitted July 2021

4.2 Open Source & Resources (refer to ELG)

4.3 Demos

Demo movie Multi Party Computation:
https://www.youtube.com/watch?v=hvBb80eXuZg