GUIDE TO PLANNING AN INFINITOME STUDY

o8t™ Infinitome

V2 2020.09

DOCUMENT VERSION 2

LEGAL DISCLAIMER

Products shown in this user guide have not been approved as medical devices or to support clinical decisions. Safety and effectiveness have not been reviewed by any regulatory agencies. o8t Infinitome is designed for research purposes only.

PATENTS PENDING

USPTO patents are pending on many elements of the processes described in this guide.

TABLE OF CONTENTS

1. INTRODUCTION

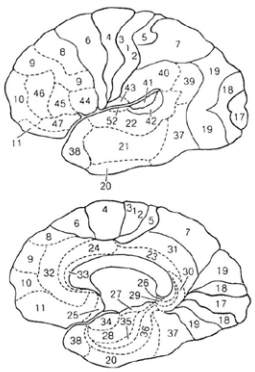
Our best research resource for studying brain diseases are the patients we treat every day. In daily practice, these patients are scanned, tested, and treated, creating billions of potential data points that are not otherwise captured in clinical research. While large sources of unrefined and unselected data may seem impractical in traditional clinical research, such sources represent invaluable commodities when using big data and machine learning methodologies.

In particular, the field of **connectomics** aims to convert big data of the brain's connections (at varying levels) into important insights on human cognition, psychology, and neuropathology.

THE CONNECTOMIC REVOLUTION

A key milestone of science in the past decade is the 'parcellation' of the human brain – that is, the optimal scaling of the cortex into components that are:

- **Sufficiently complex** to explain the origins of and changes to cognitive and psychologic function when viewed individually (single parcellations) or in combination (brain networks)
- **Sufficiently simple** that they are clinically relevant and comparable, anatomically identifiable, and do not create computationally prohibitive levels of dimensionality.



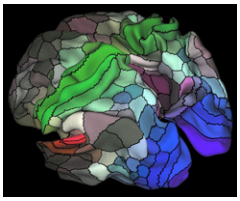


Figure 1 The Brodmann Areas (top), The HCP Parcellation Scheme (bottom)

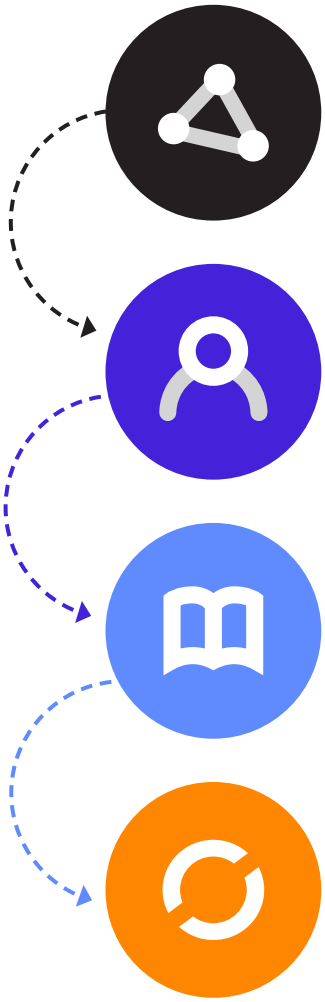
Image-based techniques to produce macro-scale parcellation models have been established through the Human Connectome Project. In total, **360 cortical regions and a further 19 subcortical regions** can be mapped. This represents a departure from classical models, such as the 52 Brodmann areas, which are less capable of explaining complex function, are inherently susceptible to subject variability, and may not provide sufficient meaningful quantitative information for statistical modeling.

FROM BREAKTHROUGH TO PRACTICALITY

Like many fields before it (e.g. genomics), connectomics has in recent years progressed through translation of core scientific breakthroughs into accessible implementations. Collective improvements in imaging technology and cloud-based software engineering have made it relatively simple for experts in clinical fields to harness the techniques of niche scientific and data fields, compared to when the HCP first began.

With appropriate pre-processing software, the creation of a 'human brain map' through a scheme of parcellations and their connective tracts is now achievable at scale, requiring only MR scans that are obtainable in routine clinical practice (a DTI and resting state fMRI); and hours, not days, of automated cloud-based processing.

MILESTONES IN CONNECTOMICS RESEARCH



The Human Connectome Project produces some of the most accurate models to data on the human brain's structural and functional connectivity.

A community of computational neuroscientists continue this work through open source software, research, and further iterations in modelling techniques.

A special 18-chapter issue of Operative Neurosurgery provides translation of these models into clinical practice.

An enterprise-grade software solutions developed to bring accessibility of connectomic frameworks to clinicians and researchers.

Figure 2 Major milestones since the commencement of the HCP in 2009

AUGMENTING EVIDENCED–BASED MEDICINE WITH DATA–DRIVEN NEUROSCIENCE

It is clear the availability of such models and the efficiency of creating them opens many doors for researchers looking to undertake deeper investigation into neurological and psychological pathology.

Arguably even more impactful, however, is that the ability to ingest and process big data of the brain enables fundamentally streamlined research workflows. Insights to support publications can be generated with a fraction of time and investment.

Using data science techniques to control for dimensionality and confounding variables, it is possible to run multiple studies and answer multiple questions on the same database. Rather than designing studies to test singles hypotheses one at a time that may lack statistically optimal origins, mining big data allows researchers to draw out the significant variables, called features, that explain phenomena of interest. Once identified, these features offer a more specific lens to view future data and may provide valuable insight into diagnosis, prognosis, and therapy.

For example, rather than a researcher forming hypothesis from observation about possible biomarkers for Alzheimer's disease, each of which would require extensive clinical study, that researcher could instead find the common connectomic features that are present in patients with Alzheimer's disease and then build models to assess future patients.

Figure 3 shows from a single database the various dimensions that a neurological disorder can be studied – with each node representing one or many potential publications. A real-world example is further provided in Figure 12.

With more quality data it is possible to have a more granular view over a specific brain disorder.

With a data science methodology, it is furthermore feasible to conduct many dimensions of study on a disorder, and publish multiple findings from a single database.

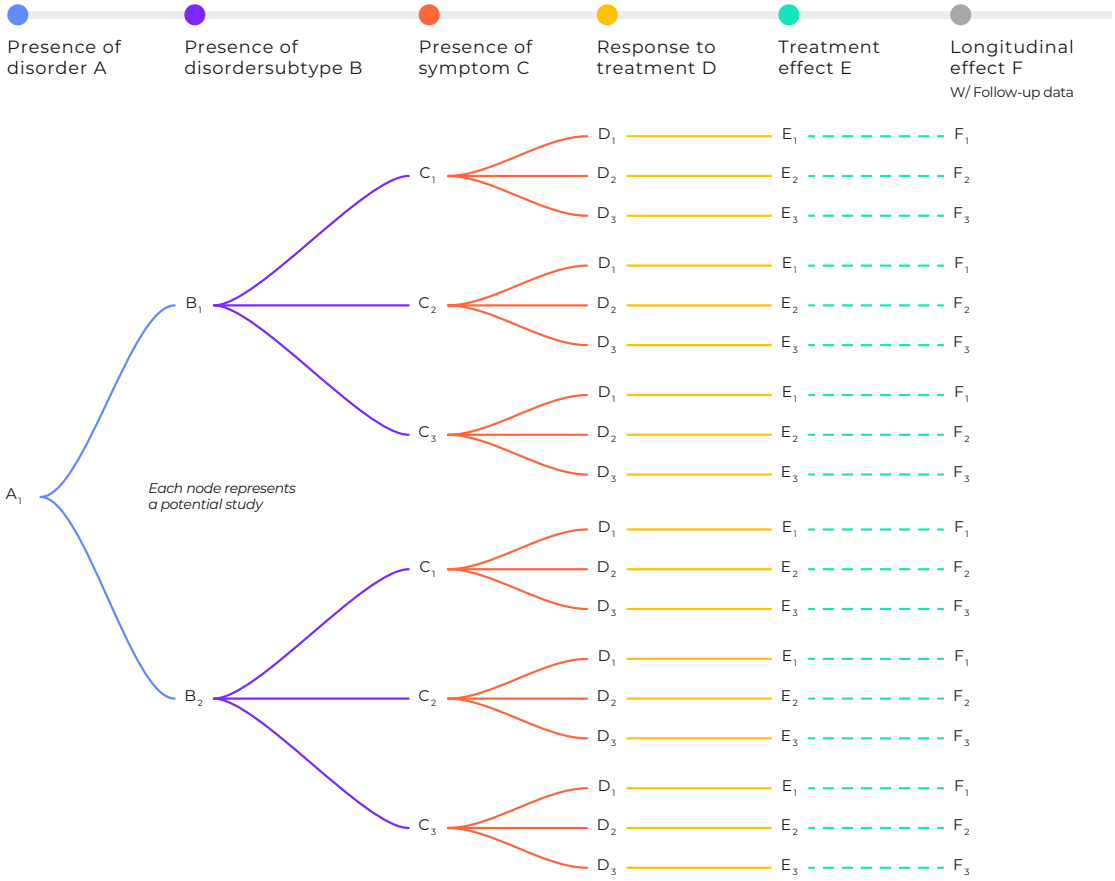


Figure 3

Understanding that such a framework is relatively novel, this guide assists clinicians and researchers to practically reframe clinical questions into machine-learning driven studies. Using powerful data processing pipelines such as **Infinitome**, clinicians and researchers can harness vast volumes of data being created by their hospitals and collaborators to deliver meaningful insights and breakthroughs that were not possible to extract previously.

Using this approach, it is possible to begin a shift to data-driven medicine. While evidence-based medicine is a gold standard, we believe that this does not preclude data-driven medicine from improving diagnosis and treatment of patients with complex neurological disorders, especially in the context of new analytical tools and methods.

2. ABOUT INFINITOME

Infinitome was developed by physicians and neuroscience researchers (with the aid of machine learning engineers, data scientists, and computer programmers), to deliver a service that is rational, clinically directed, and designed to answer the real questions posed by medical professionals.

Through a **fully-integrated cloud-based platform powered by machine learning**, Infinitome makes neuroscience research powerful and practical, enabling all centers to become precision brain medicine institutions.

All this is accomplished with minimal burden added to normal clinical workflow as it is **integrated by design**. Data collection of detailed neurologic data from consenting patients is streamlined at multiple steps. The potential of this data is then maximized through detailed yet fully automated processing.

The outputs of this pipeline can then be used to **power high-quality studies at all scales**, from single case studies to multi-institutional collaborations. In addition, use of this software can be combined with support from the Omniscient team in planning, implementation, and data analysis.

Finally, sensitive data is handled securely with HIPAA compliant protocols and state-of-the-art encrypted storage and de-identification modules. This makes it simple and legal to work with many other centers, without significant investment in infrastructure designed to handle data sharing issues.

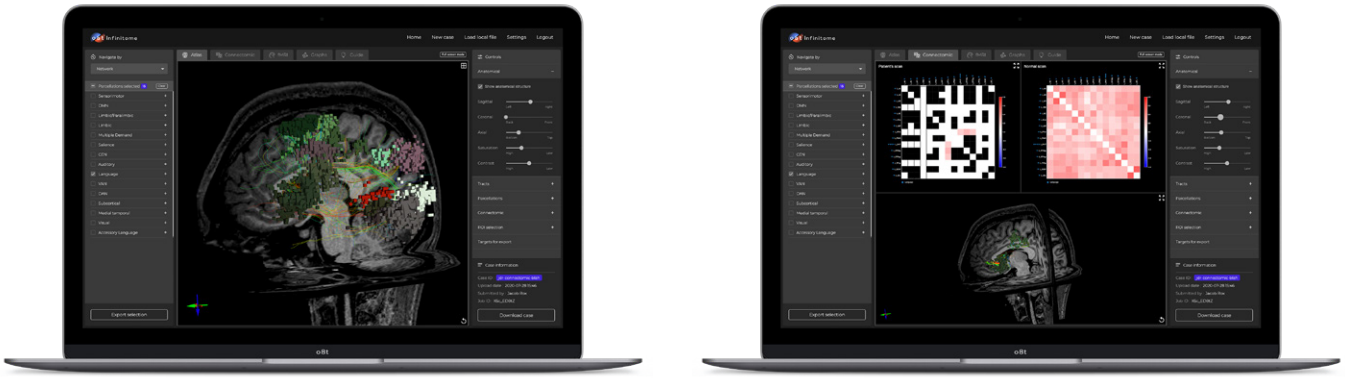


Figure 4 Screenshots of software

2.1 TOOLS FOR COLLECTION, PROCESSING, AND VIEWING

Infinitome provides a wide range of cloud-based tools to deliver researchers deeper insight with simple data entry/export workflows. The software requires images that are acquirable on commonly available MRI machines, adding only several minutes of scan time to routine scans[1], and requires no further investment in capital equipment.

AN OVERVIEW OF THESE TOOLS IS PROVIDED HERE:

Patient-specific Brain Mapping and Viewer

Figure 5 The Structural Connectivity Atlas Viewer

Through proprietary ‘re-parcellation’ techniques, our software works from the bottom-up to allow researchers to locate and visualize patient-specific networks and functional areas of the brain, including in those abnormally shaped or structurally reorganized due to injury. Maps are created using diffusion tractography data and can be processed in under an hour.

[1]Omniscient software is built to accept imaging that can be acquired in clinically relevant scan times. While results vary, it is expected that scan time should not exceed more than 20 minutes. For further information, please contact us for a full set of MR Acquisition Guidelines

Furthermore, this data can automatically be registered with other forms of patient data, which can be used to generate feature specific disease and patient models through machine learning.

Connectivity and Anomaly Assessment

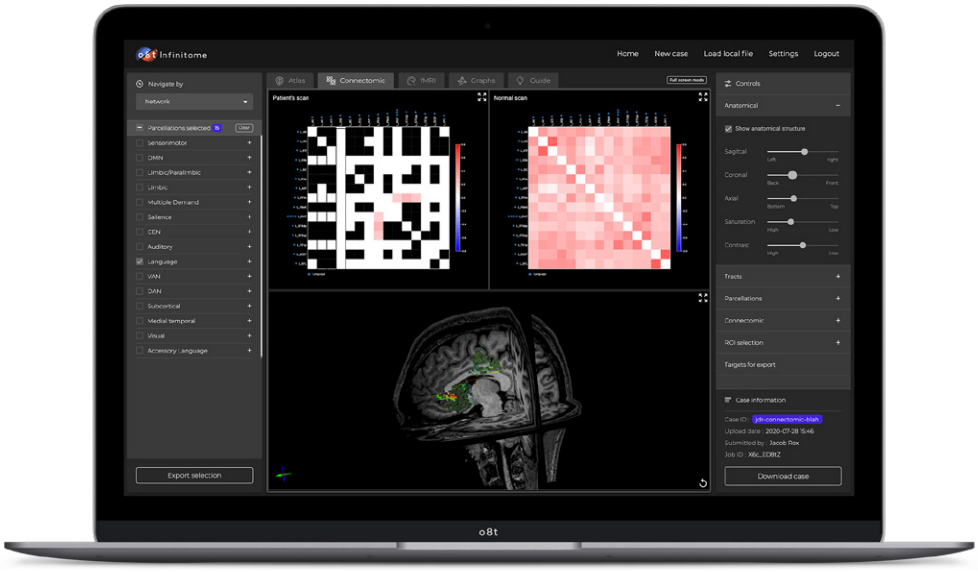


Figure 6 The Connectomic Analyses Viewer

Resting state fMRI and advanced machine learning algorithms are used to measure the connectivity between 379 subregions of the brain. Specific networks, subregions, tracts, and parcellations can be viewed together and separately, as well as compared to a large control group of healthy brains. The level of correlation between separate areas is measured, and those that are significantly and meaningfully different are identified.

Automated Data Labelling using NLP

Automated Data Labelling is a feature included in the Infinitome data collection workflow that streamlines the process of registering a scan against a patient's history and physician examination (H&P). This is accomplished through a proprietary Natural Language Processing (NLP) algorithm that can read and interpret medical records and extract relevant information.

Its inclusion eliminates a significant burden of data entry by turning medical records written in English or Chinese into large databases which can be queried using software tools. Presently, the NLP can extract salient patient data and attach these to scans from a range of type written file formats: DOC, PDF, HL7 and even JPEG of images from type-written medical records.

While medical records are limited by the documentation patterns of the writers, NLP can identify patients of a given type (for example a specific tumor type and grade) in a large dataset, identify which medications someone was on at the time of the scan, and identify confounders.

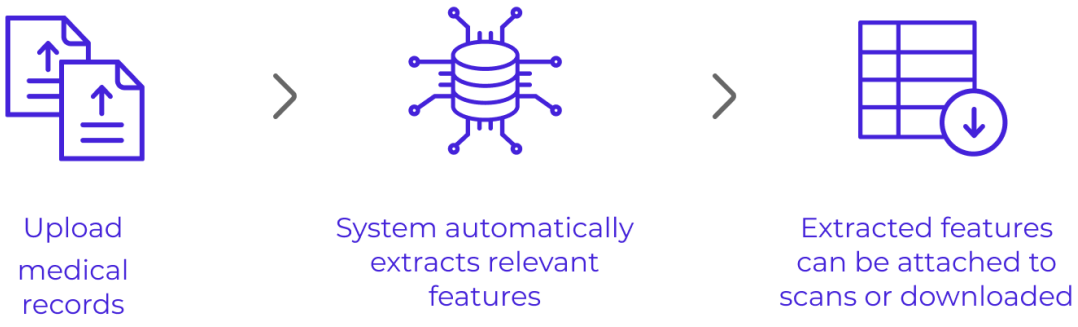


Figure 7 With Automated Feature Extraction, patient history can be rapidly processed and entered en masse in minutes, not hours.

Neurocognitive Testing Tablet App

Infinitome includes a neurocognitive testing tablet application to further streamline and improve feature labelling. The app adapts common neurocognitive tests and can be administered by trained non-experts. The benefits for expediting collection of patient data are substantial, potentially freeing up hundreds of hours of consultation from neuropsychologists.

Furthermore, being linked to the rest of the infrastructure, the app eliminates much of data entry burden by automatically registering test results to scanned patients.

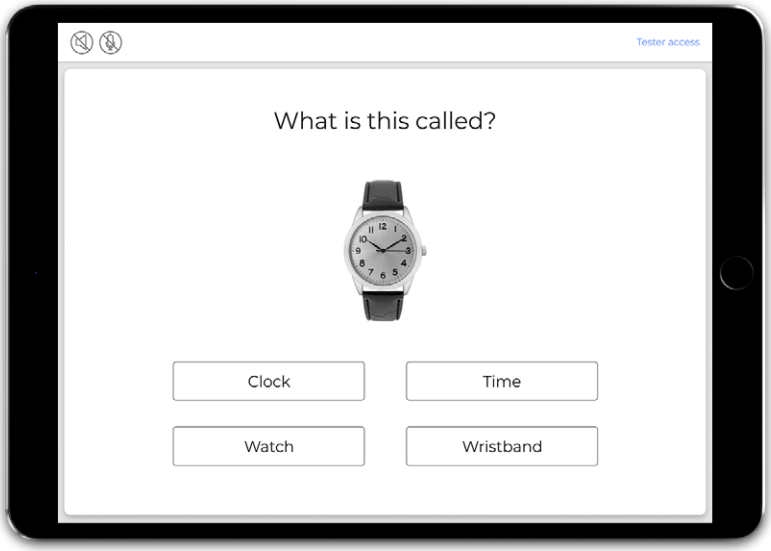


Figure 8 Neurocognitive tests can be administered through mobile devices like iPads

Current tests available on the platform include modified versions of the following:

- MMSE (general test of cognition)
- Corsi block tapping (processing and sequencing)

Planned availability for September – October 2020:

- Hopkins Verbal Learning Test (verbal memory)
- Wechsler symbol search portion (spatial memory)
- Trail making (executive function)
- Grooved pegboard test (fine motor function of hand)
- BVMTR Benton visual memory test (object memory)
- Quick Aphasia Battery (language)

Comprehensive Data Exports

While the online viewer provides visualization of processed data, Infinitome also provides the flexibility for researchers to run the outputs of their scans through other offline models. It is worth nothing that these outputs contain analyses not available on the online viewer, especially that which cannot be visualized. This exported data represents the foundation upon which post-processed machine-learning models can be produced.

There are numerous options for exporting data, including raw data, images, whole patient case files, and targets for use in image guidance equipment and for further modelling and data processing.

A full list of export files can be found in Section 5.3.

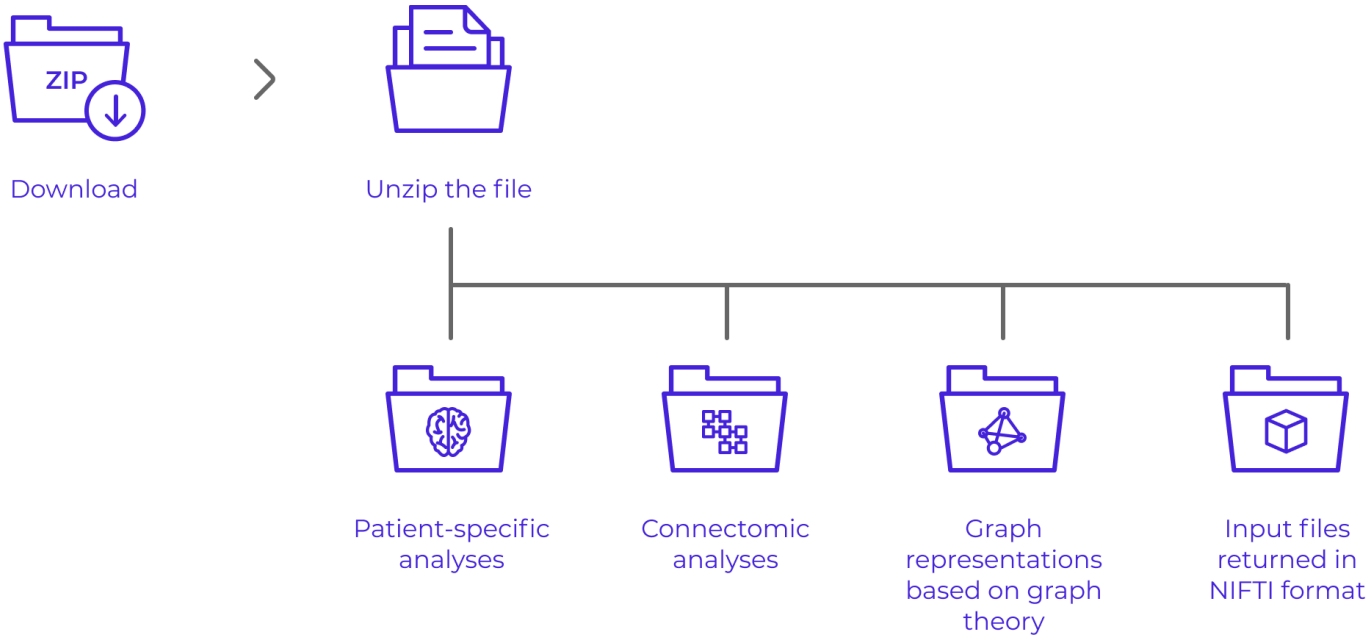


Figure 9 Overview of data export formats

3. INFINITOME: CONNECTOMICS–AS–A–SERVICE

By providing the tools discussed above, the goal of Infinitome is to allow clinicians to harness machine learning methods on their datasets using interface-based tools, i.e. acquiring and processing large sets of data without the need for experience with coding or data science. To support the design of studies such that they fully utilize the software and its outputs, consultation with our team of data scientists can be included with the service.

With this, it will be possible for you to build detailed models from your data to model and query various perspectives on a disease state, such as:

- Models for various biotypes and symptoms of a broad disorder.
- Models which learn patterns of neurological changes from different treatments.
- Models for predicting response to various treatments.

The following section provides background into the new analytical methods available to your research, and our approach to using these methods to support you. The aim is to help you better frame your clinical question into one that best utilizes these new capabilities.

WHAT IS MACHINE LEARNING?

Briefly, machine learning describes a broad field within statistics and computer science that aims to implant learning and pattern recognition capabilities, into computational systems to derive meaningful inferences from often complex and unrefined data sources. With this, computers may now be used to solve questions and make predictions that incorporate numerous data points whose connections may not be immediately clear or calculable to human observers.

Though it may be a feature of the most cutting-edge science today, machine learning, as a concept and philosophy, is neither new nor overly difficult to grasp. What is new are the advances in computation that have allowed for wider and more practical use, much of which is being used to solve modern problems, and with Infinitome to address research problems of the brain.

3.1 OUR STATISTICAL METHODS

One of the major departures of machine learning based approaches from traditional statistical analyses is the initial aim is to collect as many data points as possible, rather than to collect data relevant to an established hypothesis. This also unbounds researchers from limiting measured variables to those which are most observable, and instead finds the features, or combination of **features**, that are statistically most pertinent to the disease or function researchers are attempting to model.

As the extremely high dimensional nature of this data does not lend itself well to conventional statistical techniques, our toolset is primarily composed of machine learning model building algorithms. We can work with you to run and compare every available querying tool and machine algorithm on datasets and see what works best. In this way, machine learning is a commodity, and the goal is to ask good questions. Every symptom or response to treatment could potentially be its own model.

WHAT IS A 'FEATURE' IN MACHINE LEARNING?

In machine learning and pattern recognition, a feature is a measurable property or variable that is characteristic of the samples being observed. Its use in machine learning can be to both delineate the main driver of an outcome, or predict outcomes based on a set of drivers.

A feature can originate from any level of abstraction. For example, in a photograph, both the colors in the photo and what the resolution of the photo has been taken in represent features of the photo, with obviously different forms of information being captured.

A crucial step in machine learning is choosing or finding the feature that best explains a pattern or classification. For example, if we wanted to teach an algorithm to recognize certain species of flowers in photos, resolution may not be a very useful feature. On the other hand, color may be a very useful feature.

From there, it is worth discovering if any other feature can better be used to explain a phenomenon. This represents an important optimization step in machine learning.

Finally, for features to be usable, they must also be 'encodable', i.e. transformable into a digital format. While a computer cannot 'see' colors, it can RGB pixel values. Pertinently, connectomics has allowed far greater quantification of various 'features' of cognition and disorder.

I. Framing the Question

Careful thought about neuroanatomy, clinical observations, and relevant ways to ask a question are important to framing the data as a machine learning question. Once you have an idea of what variables you would like to study, our team of data scientists can work with you to optimize approaches for selecting the appropriate metrics and framing them as machine learning questions. The same dataset can make models for multiple questions, so thinking broadly is rewarded. We have a large variety of approaches and can try all of them.

Examples of question types relevant to clinical practice include:

- What connectomic features predict response to treatment X?
- What connectomic features relating to symptom X is present or absent within disease Y?
- How many clinically relevant subtypes of disease Y exist based on connectomic differences? What is the impact of this on treatment response or symptom clustering?
- What changes over time predict a change or progression of the disease which is clinically relevant?

II. Data Collection and Querying

The amount of connectomic data generated from the DT/fMRI scan is massive, and coupled with detailed clinical data, can provide a huge number of features from which to extract biologically relevant information into useful data. Using Infinitome, collecting and processing this kind of data at a large scale is not only possible, but relatively painless.

There are numerous methods available for querying different aspects of the connectome from the imaging and patient data – most of which were inaccessible to most researchers due to the technical difficulty implementing analytical tools without specialized expertise.

THESE INCLUDE:

- **Structural Connectivity Matrices**
  - Measuring the number of structural connections between parcellations from tractography
  - Over 143,000 features
- **Functional Connectivity Matrices**
  - Measuring correlative connectivity between parcellations from resting state fMRI data
  - Over 143,000 features
- **Graph Metrics**
  - Global Efficiency
  - Eigenvector Centrality
  - Modularity
  - Segregation
- **Focal Diffusion Metrics**
  - Fractional Anisotropy
  - Dynamic Functional Connectivity
  - Dynamic Graphs
  - Sliding Windows
  - Structure Flow on Manifolds

Figure 10 A full connectivity matrix representing 143,300 data points and potential features

- **Hierarchy and Cortical Gradient Analysis**
- **Measures of Causality**
- **Patient Data**
  - Known Confounders
  - Demographic
  - Disease States
  - Medical History
  - Clinical Grading Scales
- **Clinical Grading Scales**
- **Neurocognitive Features**

III. Sample Size and Power Determination – Number needed per study

To our knowledge, there is no gold standard in literature about methods for performing power calculations for predicting the sample needed to achieve a specific AUC for a machine learning model. This is especially true with deep learning methods which contain a stochastic component in their methods for re-weighting features in the hidden layers to minimize the cost function. Furthermore, when transitioning from an analysis of disease to symptom, the degree of variance of symptoms is unclear, adding further uncertainty to what effect size is being modelled when we estimate the number needed to create a model with AUC>0.95 (the generally accepted ideal model goal for accuracy).

As a method for addressing this, our data scientists created a novel approach which estimates the effect size using existing datasets on previously trained machine learning models on functional and structural connectivity matrices which had achieved AUC=0.7 in models of schizophrenia and autism. They generated an extrapolation of this estimate of the sample size-AUC curve with a power law distributed modeling of diminishing rate of returns with increasing numbers assuming normal distribution of errors. Based on this, we estimate that to generate a model with AUC>0.95 based on brain connectivity data from brain

images acquired using the parameters outlined above, it may require between 1000-2000 subjects to form definitive answers on a host of clinically relevant question.

Given this assessment, and the fact that we are studying potentially numerous sub-questions within a study, it is obvious that the patient numbers needed to create models of this quality which address all the possible questions one can raise in this area are substantial. This is the principle motivation for engaging in a long term, massive scale project. While we do not anticipate being able to achieve models of this rigor on all possible questions, we do think that this justifies enrolling as many patients as possible.

IV. Addressing Bias and Confounding

Addressing bias and confounding is a key challenge of conventional clinical study design. Strict patient selection criteria can lead to delays in enrolment and progression of the study, while selection criteria based on inaccurate assumptions can lead to inconclusions or false conclusions.

On the other hand, making available larger and higher dimensional datasets enables the opportunity to use better sampling methods and strategies to identify and deconvolute confounders at a patient level, normalizing them before comparing with others. Using higher dimensional data, however, necessitates more advanced techniques derived from data science fields: firstly, through **stratification** and then applying **machine learning approaches**.

By stratifying the data, you can ask focused questions that subset large data sets into several key points and areas of interest, such as by different parts of the brain, functional networks, and disease grades.

It is worth noting the potential challenges of this step:

- For certain variables, especially those that are symptom based, it may be challenging to set appropriate cut-offs.
- It is also important that that different stratifications are weighted equally.

For these challenges, our data science team can provide assistance.

Within the data subsets, you can then apply various techniques to model and classify both linear and non-linear models. Our team has worked extensively with these models and can provide assistance in applying approaches discussed in the following section.

3.2 OMNISCIENT'S MACHINE LEARNING PIPELINE

Once data is uploaded and processed, and the research questions have been framed as federated machine learning questions, you can begin applying machine learning algorithms to build predictive models.

The following section serves as a primer commonly used techniques. When partnering with researchers. our workflow is comprehensive and ensures diligence when forming the machine learning methodology. We will try all approaches, see what fits, then readjust accordingly. Dimensionality reduction is essential for addressing extremely complex data sets like this. In addition, biologically relevant methods are key to reducing the effect of scanner parameters and individual differences on overall data analysis.

Baseline and demographic characteristics (age, gender, disease, treatment etc) will be studied to determine if they interact with the datapoint of interest using dimensionality reduction techniques and other approaches described in the next section. For purposes of data presentation for publication, they will be described as continuous variables or proportions as needed.

WHAT IS A DIMENSIONALITY REDUCTION?

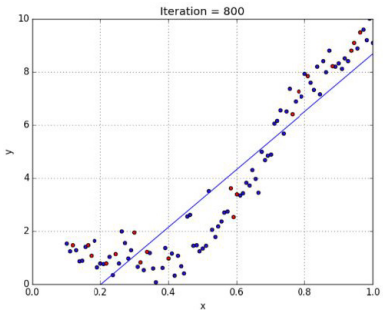
Dimensionality reduction is the transformation of data from a high dimensional space to a low dimensional one, while aiming to retain the meaningful properties of the original data. Data of excessive complexity may not be useful for a machine learning algorithm just as it may not be to a human.

Machine Learning Modelling Approaches

The following are a list of some of the standard modeling approaches that will be used:

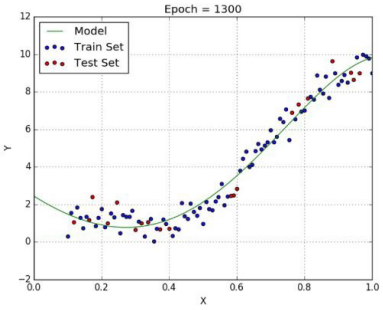
Linear Regression

Linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).



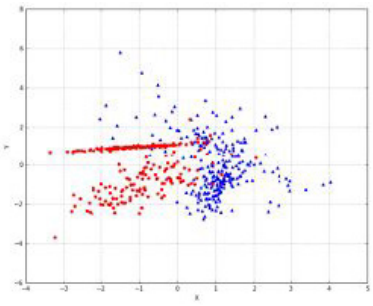
Polynomial Regression

Used to describe non-linear phenomenon, such as growth rate of tissues or disease epidemics. Relationship between the independent variable x and the dependent variable y is modelled as an n th degree polynomial in x .



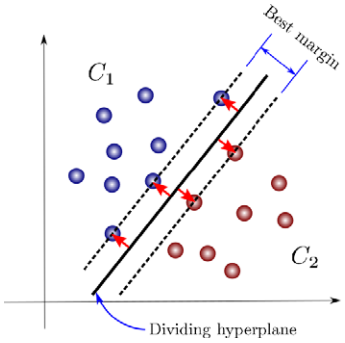
Naive Bayes

Family of algorithms that uses probability theory and statistical independence to classify data. All data points are independent of each other, and the probability of an event is adjusted as new data is introduced.



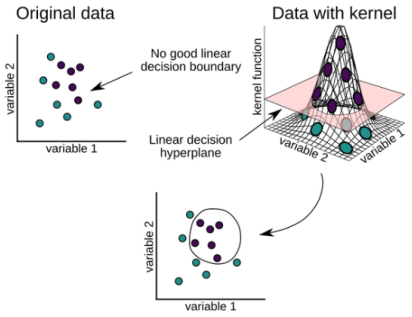
Support Vector Machine (SVM)

Performs classification and regressionanalysis, by sorting data into categories. With labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples. It outputs a map of the sorted data with margins as far as possible.



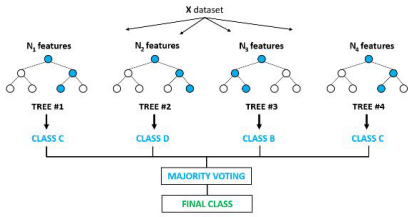
Kernel SVM

Uses pattern analysis to find and study general types of relations, such as correlations. The kernel function allows for separability of non-linear regions, by placing a two-dimensional plane into a higher dimensional space.



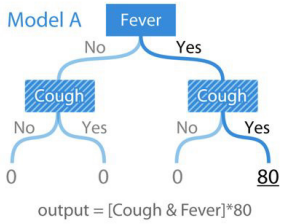
Random Forests

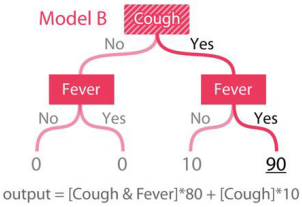
Classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.



Gradient Boosting (such as XGBoost)

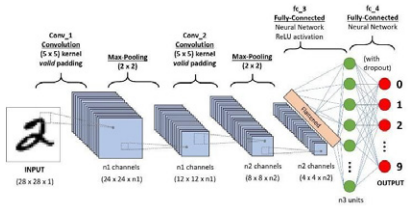
Decision-tree based algorithm used for classification and regression predictive modeling. New models are created that predict the residuals or errors of prior models and then added together to make the final prediction.





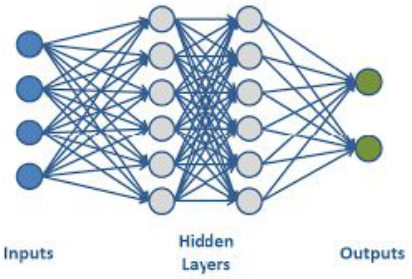
Convolutional Neural Network (CNN)

Deep learning algorithm for image classification. Takes in an input image, assigns importance to various aspects, and differentiates one from another. Images are processed as matrices of numbers with additional dimensions.



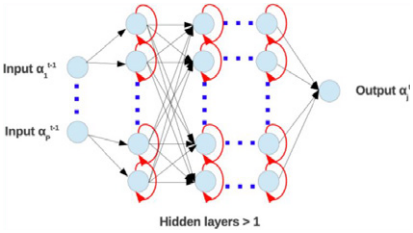
Recurrent Neural Network (RNN)

Deep learning algorithm that recognizes a data's sequential characteristics and uses patterns to predict the next likely scenario Uses feedback loops and the output from the previous step is fed as an input to the current step.



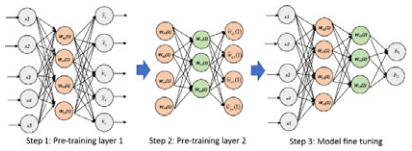
Long Short-Term Memory (LSTM)

Type of RNN deep learning algorithm capable of learning long-term dependencies, which is useful for certain types of predictions. It recognizes patterns in sequences of data by taking time and sequence into account.



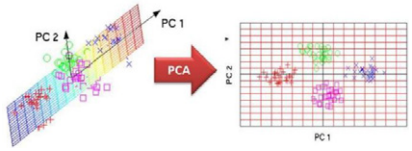
Stacked Autoencoder

Used for dimensionality reduction by training the network to ignore signal noise. Uses several layers of spared auto encoders, where output of each hidden layer is connected to the input of the successive hidden layer.



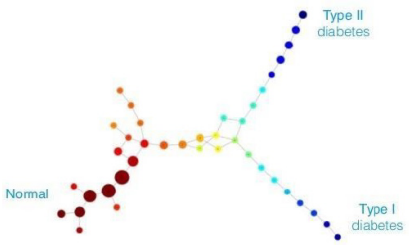
Principal Component Analysis (PCA)

Reduces the dimensionality of a dataset consisting of many variables correlated with each other, while retaining the variation present in the dataset. This is one by transforming the variables into a new set of variables, ordered such a way as the variation present decreases as it moves down the order.



Topological Data Analysis

Applies the tools of algebraic topology and statistical learning to find a quantitative structures in the dataset and reduce noise. Allows for persistent homology, which can detect small or large features called holes, which is useful in computation biology.



4. INFINITOME IN ACTION

4.1 PLANNING A RETROSPECTIVE STUDY

While working with Omniscient on a prospective study will ensure that the best collection protocols are observed, you can build models from existing data provided they meet MRI scan requirements and can be paired to clinical notes.

MRI Scan Requirements

- T1 high resolution
- DTI with 30 directions or more
- Resting state fMRI

Clinical Data

This can be medical history, clinical, outcome scales, and/or neurocognitive data. These are treated as “feature labels” for machine learning analysis and displays what was going on with the patient at the time of the scan. In terms of research objectives, these can be either outcomes of interest, demographics of interest, or confounders you want to control for.

Analysis & Output

This data can then be run through a battery of machine learning algorithms and tests, customized based on your research objectives. The correlation between all regions of the patient’s brain is examined and can be compared to different types of control groups. Individual parcellations or entire networks can be studied. Data can be provided in multiple forms, including raw, matrices, and graphs.

4.2 PLANNING A PROSPECTIVE STUDY

The ideal structure of a prospective study is to capture pre- and post- treatment data, which would allow for studying changes in the connectome, and how different subsets respond to treatments.

Before undertaking a prospective study, you can build a retrospective study on existing data, treating it as a pilot study from which you can test concepts, build hypotheses, and formulate a more sophisticated structure for a future prospective study.

The figure on the right shows a visual example of the type of output generated from a retrospective study using Infinitome. This study consisted of 82 dementia patients, represented in the columns. All 379 parcellations of their brains are represented in the rows with colors describing centrality measures.

Green squares represent the highest ranked parcellations and the red squares are the lowest ranked parcellations in each individual. Note that by holding the order fixed to the overall rankings for the group, we provide a simple way to demonstrate variations in hubness, and highlight when centrality drops, as with Alzheimer's disease which preferentially attacks the hub regions.

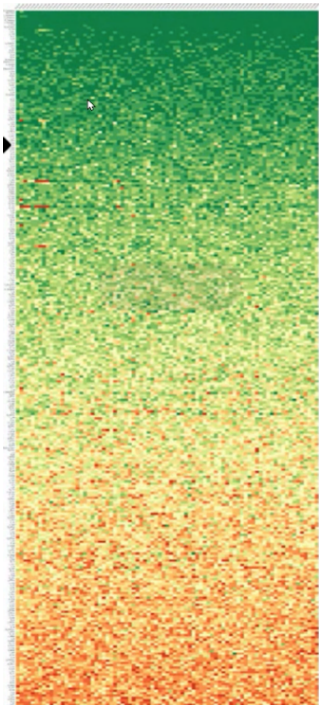


Figure 11

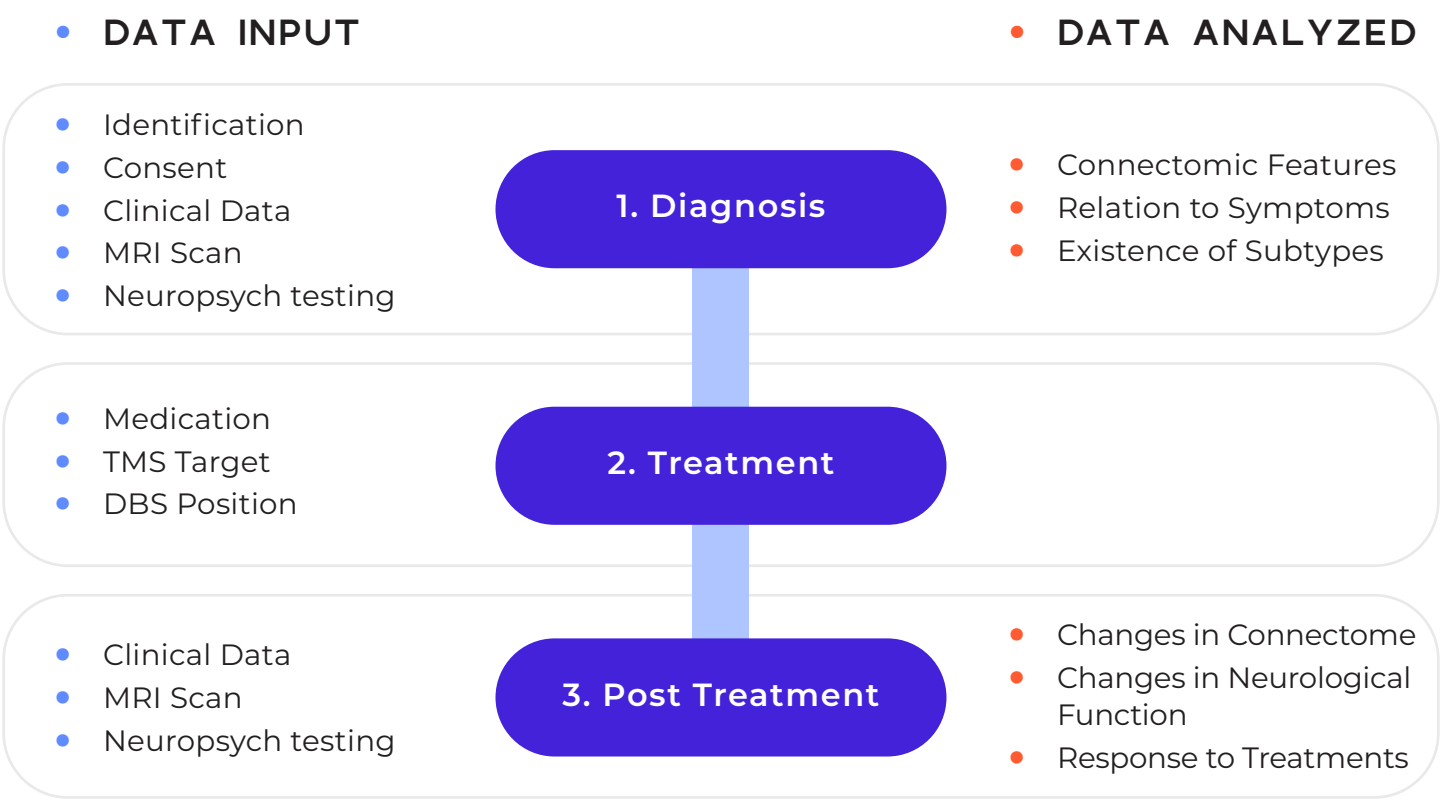
MRI Scan

- T1 / DTI / resting state fMRI
  - Before treatment
  - After treatment

Clinical Data

- TMS target or electrode position (if applicable)
- Outcome scales:
  - All subsets recorded separately
  - Pre and post treatment
- Past medical history
- Medication history
- Neurocognitive testing
- Psychiatric scales

Basic Structure



4.3 EXAMPLE OF A MULTI–CENTER PROSPECTIVE STUDY: THE GLIOMA CONNECTOME PROJECT

GLIOMA
CONNECTOME PROJECT

STUDY OVERVIEW

This is a consortium of leading brain tumor centers dedicated to exploring the clinical applications and implications of understanding the structural connectome of the human brain as it shapes: a) the natural history of glioma growth; b) surgical planning; c) predicts response to surgical, radiation, medical and immunological therapies, with the objective of changing current treatment paradigms, developing new biomarkers and endpoints, measuring the benefits and risks of glioma therapy including surgery, radiation, chemotherapy, electrical field therapy, and others.

DATA INPUT		DATA ANALYZED
• Identification • Consent • Clinical Data • MRI Scan • Neuropsych testing	**1. Diagnosis**	• History & Physical • Neurological Function • Quality of Life
• Intraoperative Data • Clinical Data • Neuropsych testing	**2. Surgery**	• Pre-op Connectome • Tumor Anatomy • Intraoperative Stimulation Points
• Clinical Data • MRI Scan • Neuropsych testing	**3. Post-Operation**	• Post-op Connectome • Post-op Neurological Function • Area Resected
• Clinical Data • MRI Scan • Neuropsych testing	**4. 1st Followup**	• Connectome Neurological Function • Tumor Location
• Clinical Data • MRI Scan • Neuropsych testing	**5. Routine Followup**	
• Clinical Data • MRI Scan • Neuropsych testing	**6. Recurrence Surgery**	• Pre-op Connectome • Tumor Anatomy • Intraoperative Stimulation Points • History & Physical • Pre-op Neurological Function • Quality of Life
• Clinical Data • MRI Scan • Neuropsych testing	**7. Post Operation 2**	• Post-op Connectome • Post-op Neurological Function • Area Resected

STRUCTURE OF STUDY

Analysis of Primary Outcomes of Interest

- To address the numerous potential questions inside the primary outcome of interest, we will frame the machine learning question in the following way:
- First, we will pose the question in terms of a neurocognitive question; an example of this is “What connectomic changes predict anomic aphasia after surgery”.
- We will identify the neurocognitive feature of interest and look for interdependencies within this feature and other neurocognitive tests.
- We will identify patients who had this neurologic problem after surgery who did not have it before using cloud-based querying tools and create a dataset of those patients who did not have these deficits as controls.
- We will then perform data processing approaches using algorithms to determine maps of functional and structural differences between the preoperative and postoperative images.
- We will use dimensionality reduction techniques to determine possible in the clinical or neurocognitive data which are confounding variables worth controlling for.
- Machine learning batteries described above will be used to fit a model which predicts the development of this deficit using connectomics difference maps with the confounding variables and latent space components fit into the model.
- If necessary, black box analyses will be used to determine the relative importance of various connectomics features which will provide insight into the relative importance of connectomics changes to the development of neurocognitive changes.

4.4 EXAMPLE OF SINGLE CENTER PROSPECTIVE STUDY: A SCHIZOPHRENIA STUDY

As a demonstrative example, schizophrenia is multifaceted in both presentation and therapy, with each facet an opportunity to build detailed models to further our understanding of the disorder. These models can be built for both broader and more granular clinical questions.

DATA COLLECTION

At a minimum, Infinitome requires the MRI scans outlined below:

- T1, DWI, resting state fMRI
  - Before treatment
  - After treatment

Then, depending on the area of interest, the following may be potential clinical data of interest.

- Past medical history
- Medication history
- Positive and Negative Syndrome Scale (PANSS)
- With all subsets recorded separately
- Pre and post treatment
- TMS target
- Neurocognitive testing (completed through Infinitome App):
  - MMSE
  - Hopkins Verbal Learning Test
  - Wechsler symbol search portion
  - Trail making
  - Grooved pegboard test
  - BVMTR Benton visual memory test
  - Corsi block tapping
  - Quick Aphasia Battery
- Psychiatric scales:
  - HAMD
  - General anxiety rating scaling

ANALYSIS & OUTPUT

With the detailed data generated from a large prospective study, it may be possible to generate several detailed models.

These models may be able to address the following:

- Presence of Disorder
  - Schizophrenia
  - Schizo Affective
  - Absence
- Specific Symptoms
  - Over 30, taken from PANNS
  - Includes delusions, disorganization, auditory hallucinations, visual hallucinations, and hostility
- Connectivity Features
  - Strong or weak correlation between parcellations
  - Compared to normal control group
- Response to Treatments
  - TMS targets
  - Medications, such as CBT, D2 blockers, and anti-psychotics
  - Initiating treatment and increased dosage
- Changes in Connectivity
  - How connectome patterns change from treatment
  - Structural connectivity from diffusion tractography
  - Functional connectivity from fMRI

POTENTIAL MODELS GENERATED (EACH NODE CAN REPRESENT AN INDIVIDUAL STUDY):

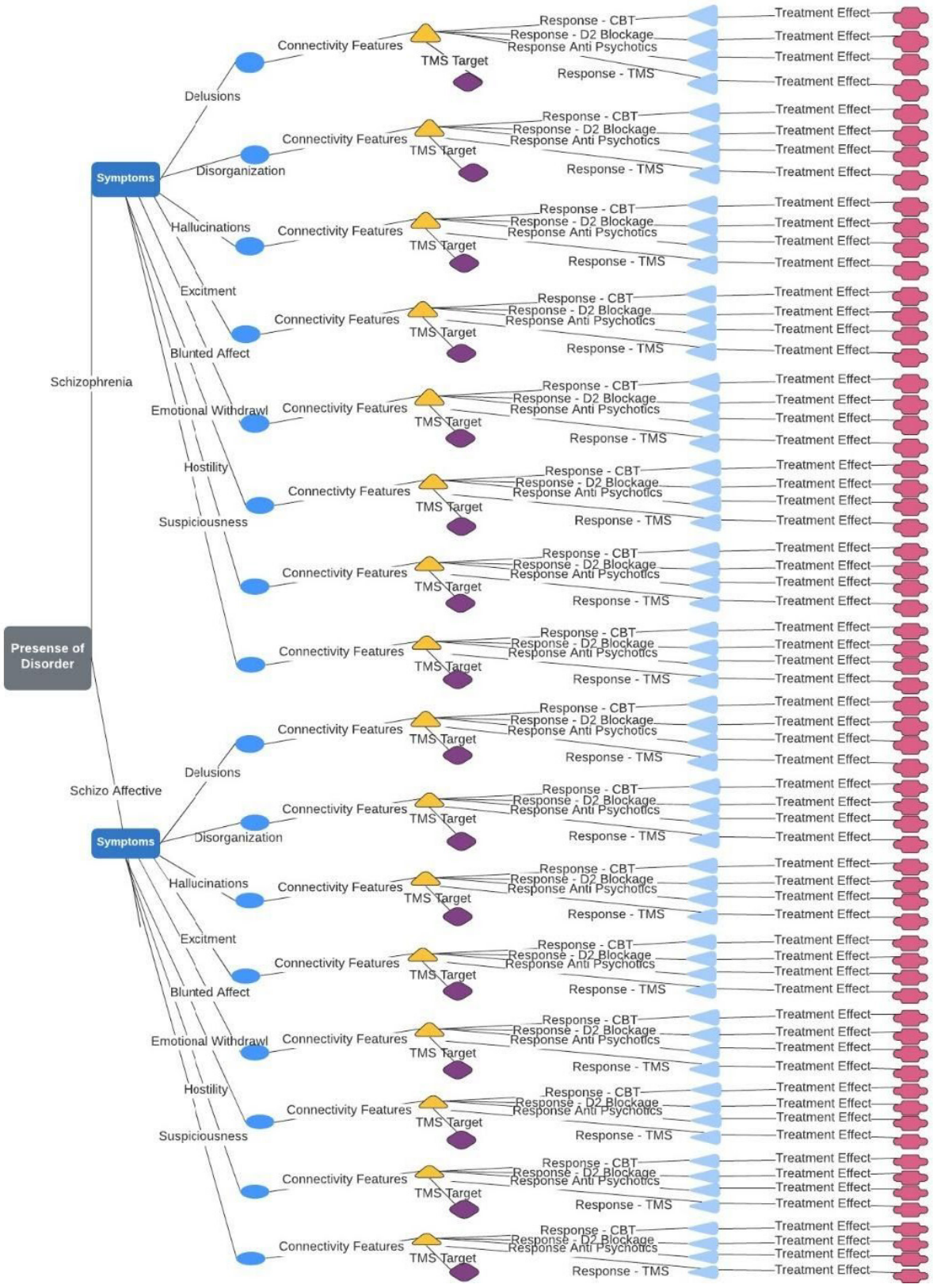


Figure 12

5. DETAILED PRODUCT INFORMATION

5.1 MRI ACQUISITION PROTOCOL

Infinitome and all Omniscient software is designed to work with commonly available MRI scanners and add only minimal additional scan time.

Below is an example of a routine MR protocol. A full guide can be made available upon request.

SERIES	SCAN	PARAMETERS
1	Scout	
2	Calibration	
3	dMRI	2 mm x 2 mm x 2 mm voxels • FOV = 25.6 cm • Matrix = 128 mm x 128 mm • Slice thickness = 2.0 mm b = 0, b = 1000 40 directions, bipolar/no gradient overplus Gap = 0.0 mm TE = Shortest, TR = Shortest Flip angle = 90 Full-brain coverage

Continue on next page.

SERIES	SCAN	PARAMETERS
4	rs-fMRI	3 mm x 3 mm x 3 mm voxels • FOV = 240 cm • Matrix = 80 mm x 80 mm • Slice thickness = 3.0 mm 180 volumes/run TE = 30 ms TR = Shortest Flip angle = 90
5	Axial T2 FLAIR (only if requested)	FOV = 25.6 cm Matrix = 256 x 256 Slice thickness = 2.0 mm (2D) Gap = 0.0 Full brain coverage
6	Axial 3D T1 pre-contrast (only if requested)	3D FFE FOV = 25.6 cm Matrix = 256 x 256 Slice thickness = 1.0 mm (3D) Axial Gap = 0.0 TE = Shortest TR = Shortest Full head coverage
7	Contrast Injection	
8	Axial 3D T1 post-contrast	3D FFE FOV = 25.6 cm Matrix = 256 x 256 Slice thickness = 1.0 mm (3D) Axial Gap = 0.0 TE = Shortest TR = Shortest Full head coverage

5.2 MRI PROCESSING METHODS

Overview

Images uploaded to Infinitome will be preprocessed using our proprietary cloud-based image processing software, which creates a subject specific version of the Human Connectome Project's Multimodal Parcellation (HCP-MMP1) atlas using diffusion tractography (DT). It also performs analytics on both DT and resting state functional MRI (rsfMRI) data. Images are not warped to any standard imaging spaces (MNI, Tailarch, etc), but rather structural connectivity is used to locate and map out all the individual parcellations based on the patient's native space, making it truly personalized. All processing steps are performed in the Python programming language.

Several steps below are currently have USPTO patents pending.

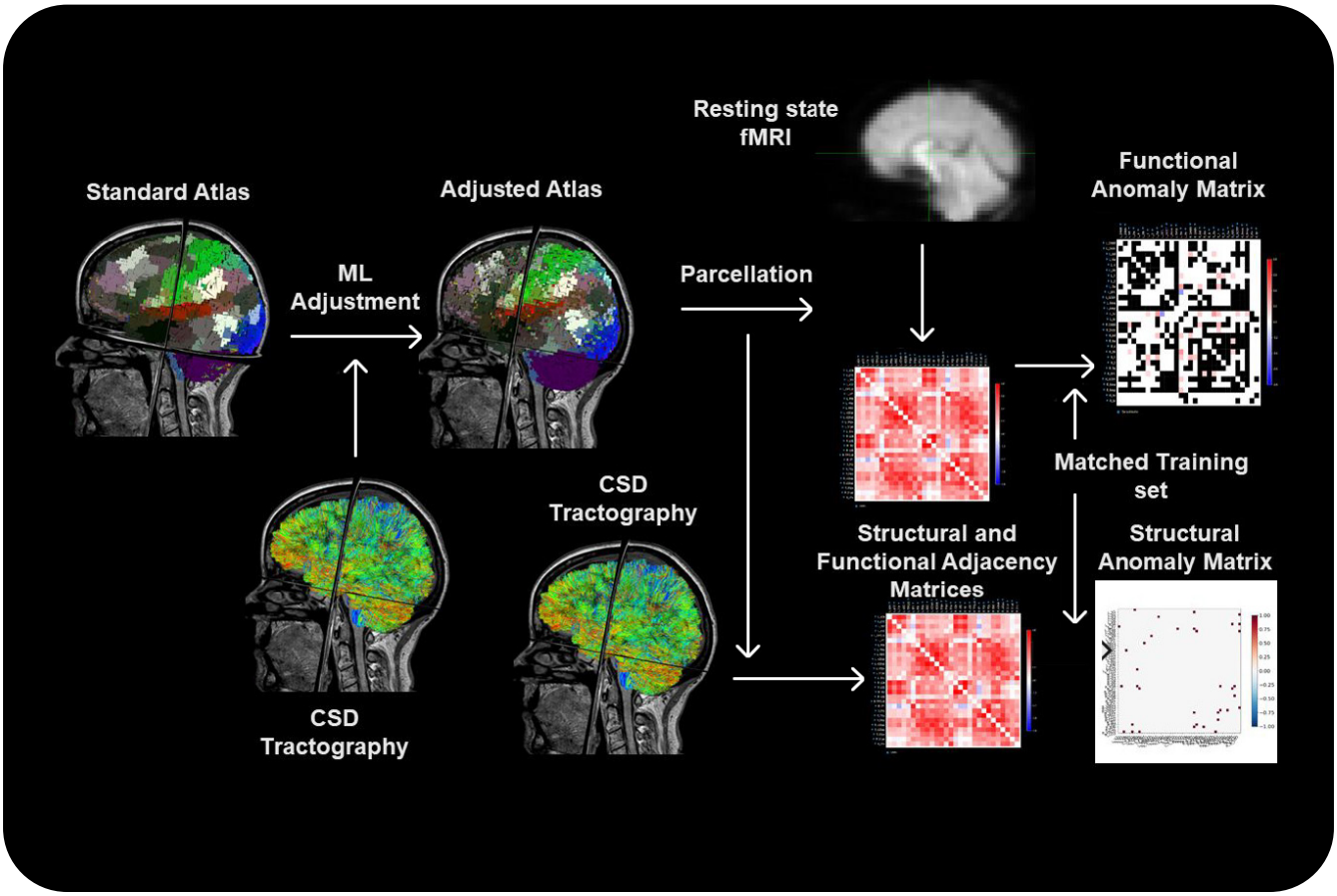


Figure 13 Schematic of pre-processing pipeline

I. Diffusion Tractography Pre-processing Steps

The DT images are processed using standard processing steps which specifically include the following procedures:

1. Diffusion image is resliced to ensure isotropic voxels.
2. Motion correction is performed using a rigid body alignment.
3. Slices with excess movement (defined as DVARS> 2 sigma from the mean slice) are eliminated.
4. T1 image is skull stripped using a convolutional neural net (CNN), this is inverted and aligned to the DT image using a rigid alignment, which is then used as a mask to skull strip the DT.
5. Gradient distortion correction is performed using a diffeomorphic warping method, which aims to locally similarize the DT and T1 images.
6. Eddy current correction is performed.
7. Fiber response function is estimated and the diffusion tensors are calculated using constrained spherical deconvolution.
8. Deterministic tractography is performed with random seeding, usually creating about 300,000 streamlines per brain.

II. Creation of a Personalized Brain Atlas using Machine Learning

Our software creates a machine learning based, subject specific version of the HCP-MMP1 atlas based on diffusion tractography structural connectivity. This novel method was created by training a machine learning model on 200 normal subjects by first processing T1 and DT images as above. A HCP-MMP1 atlas in NIFTI MNI space is then warped onto each brain and the structural connectivity calculated between every pair of this atlas and a set of ROI containing 8 subcortical structures per hemisphere and the brainstem based on the streamlines which terminated within an ROI. These feature vectors for each region were then used as a training set and the data were modeled using the XGBoost method.

This model is then applied to the new subject by first warping the HCP-MMP1 atlas to the new brain and collecting a set of feature vectors of the connectivity of each voxel. The feature vectors are then used to determine if each voxel belongs to a parcellation or region or not, and if so to assign the voxel to that parcellation. This creates a version of the HCP-MMP1 atlas (with subcortical components, which is not dependent on brain shape or pathologic distortion, and which is specific for this subject, but comparable between subjects.

III. Resting State fMRI Pre-processing Steps

The rsfMRI images are processed using standard processing procedures which specifically include the following steps:

1. Motion correction is performed on the T1 and BOLD images using a rigid body alignment.
2. Slices with excess movement (defined as DVARS> 2 sigma from the mean slice) are eliminated.
3. T1 image is skull stripped using a convolutional neural net (CNN), this is inverted and aligned to the resting state bold image using a rigid alignment, which is then used as a mask to skull strip the rsfMRI image.
4. Slice time correction is performed.
5. Global intensity normalization is performed.
6. Gradient distortion correction is performed using a diffeomorphic warping method, which aims to locally make the rsfMRI and T1 images geometrically similar.
7. High variance confounds are calculated using the CompCor method; these confounds as well as motion confounds are regressed out of the rsfMRI image, and the linear and quadratic signals are detrended. Note this method does not perform global signal regression.
8. Spatial smoothing is performed using a 4mm FWHM Gaussian kernel.

IV. rsfMRI Correlation and Outlier Detection

The personalized atlas created in previous steps is registered to the T1 image and localized to the grey matter regions. Thus, it is ideally positioned for extracting an average BOLD time series from all 379 areas (180 parcellations x 2 hemispheres, plus 19 subcortical structures). This yields 143,641 correlations.

Outlier detection using a tangent space connectivity matrix is performed by comparing the results with a subset of 200 normal subject resting state fMRI samples in whom a tangent space connectivity transformation is performed to determine the range of normal correlations for each functional connectivity pair in the matrix. Abnormal connectivity is determined as a 3-sigma outlier for that correlation, after excluding the highest variance 1/3 of pairs, to further reduce the false discovery rate. Assignment of parcellations to various large-scale brain networks is based on several previous coordinate based meta-analysis and matching the HCP-MMP1 parcellations to the coordinates of the ALE in MNI space, which has been previously published (or in review presently) by our group.

5.3 DATA EXPORT PACKAGE

Our downloadable export makes it easy to move data back into your existing workflows. The downloadable export package contains scan results in your preferred programming language and format and goes into further detail than what is available on the online viewer.

A full list of current export files is provided below:

EXPORT FILE NAME	DESCRIPTION	RESEARCH APPLICATIONS
/T1_coronal_thumbnail.png'	MRI Reference Image - Coronal	Anatomical 2-dimensional background images for visual representation of data
/T1_axial_thumbnail.png'	MRI Reference Image - Axial	
/T1_sagittal_thumbnail.png'	MRI Reference Image - Sagittal	
/Connectomic/bold_img clean.nii.gz'	Pre-processed functional MRI (fMRI) BOLD weighted image	3-dimensional functional image for visual representation of data
/Connectomic/fitted/ laterality_connectome fitted.csv'	Laterality scores - Fitted Atlas	Scores for determining if the network, pathway or pathology being investigated is neurally asymmetrical
/Connectomic/ conservative/ laterality_connectome_ conservative.csv'	Laterality scores - Structural connectivity atlas	
/Connectomic/fitted/ reindex_output_fitted_ full_ordered_csv.zip'	Functional connectomics - Overall + network level - Fitted atlas	Raw connectomic values between all parcellations (379 x 379)
/Connectomic/ conservative/reindex_ output_conservative_full_ ordered_csv.zip	Functional connectomics - Overall + network level - Structural connectivity atlas	

Continue on next page.

EXPORT FILE NAME	DESCRIPTION	RESEARCH APPLICATIONS
/Connectomic/fitted/ reindex_output_fitted.zip'	Functional connectomics - Overall + network level - Fitted atlas	Connectivity matrix displaying anomalous connectivity values within networks of interest
/Connectomic/ conservative/reindex_ output_conservative.zip'	Functional connectomics - Overall + network level - Structural connectivity atlas	
/Connectomic/fitted/ classifier_df_fitted.csv'	Mental illness classification scores - Fitted Atlas	
/Connectomic/ conservative/classifier_df_ conservative.csv'	Mental illness classification scores - Structural connectivity atlas	
/Input/input_dwi.nii.gz'	Inputted DWI	Extraction of raw input image files for future use
/Input/input_dwi.nii'	Inputted DWI	
/Input/input_bvec.bvec'	Inputted DWI gradient direction values	
/Input/input_bold.nii.gz'	Inputted BOLD	
/Input/input_t1.nii.gz'	T1 Anatomical Input	
/Input/input_bval.bval'	DWI field strength values	

Continue on next page.

EXPORT FILE NAME	DESCRIPTION	RESEARCH APPLICATIONS
/Graph/graph_output_zip.zip'	Graph metrics - global efficiency	Vairety of metrics for parcellations and tracts, can be applied to a number of research interests
/Atlas/atlas_reindex_bundle.zip'	NIFTI files for every selected parcellation	Overlay creation of individual parcellations onto a background image
/Atlas/T1.nii.gz'	T1 Anatomical Scan (Skull stripped)	3-dimensional anatomical atlas for future overlay of affiliated 3-dimensional image sets
/Atlas/T1.nii'	T1 Anatomical Scan (Skull stripped)	
/Atlas/t1_tracts.trk'	Complete tractography analysis file	Reintegration into image analysis pipelines (.trk file)
/Atlas/t1_tracts_mini.trk'	Portioned tractography analysis file, initial upload prior to full segmentation and CSD calculation	
/Atlas/liberal/t1_atlas_liberal.nii.gz'	Structural connectivity atlas	3-dimensional representation of tracts between functional parcellations for overlay use. (NIFTI format)
/Atlas/liberal/parcellations_liberal.vtp'	Structural connectivity atlas	
/Atlas/liberal/missing_parcellations_output_liberal.csv'	List of missing parcellations in structural connectivity atlas	Identifying missing tracts unable to be identified by the ML model, potentially due to pathology (eg. Tumor)

Continue on next page.

EXPORT FILE NAME	DESCRIPTION	RESEARCH APPLICATIONS
/Atlas/conservative/missing_parcellations_output_conservative.csv'	List of missing parcellations in conservative atlas	Identifying missing parcellations unable to be identified by the ML model, potentially due to pathology (e.g. Tumor)
/Atlas/fitted/missing_parcellations_output_fitted.csv'	List of missing parcellations in fitted atlas	
/Atlas/fitted/t1_atlas_fitted.nii.gz'	All displayed parcellations - fitted atlas	
/Atlas/conservative/t1_atlas_conservative.nii.gz'	All displayed parcellations - conservative atlas	3-dimensional representation of functional brain parcellations for overlay use. (NIFTI format)
/Atlas/fitted/parcellations_fitted.vtp'	Fitted atlas	
/Atlas/conservative/parcellations_conservative.vtp'	Conservative atlas	

5.4 DATA SECURITY

I. Overview

Infinitome adheres to all relevant medical data and cybersecurity laws of the country. Cloud servers are located in various jurisdictions and no raw patient data will leave the country of origin.

Infinitome uses advanced cloud-based technology for transmitting, processing, and storing brain imaging and medical record data with minimal risk to PHI, and complies with HIPAA and GDPR standards. Depending on implementation methodology, PHI is stripped from DICOM headers prior to processing either in our cloud servers (manual uploads, Figure 14), or in a gap server (PACs integration, Figure 15). Note that this means that no PHI is processed or stored by the system.

The datasets from different time points for a single patient are clustered together by generating a unique encryption-based mapping using the MRN or a user value to create a unique system identifier.

The following diagrams provide an overview of the Omniscient IT and Security process.

SECURITY SCHEMATIC

Data Transmission - Chrome Browser Option

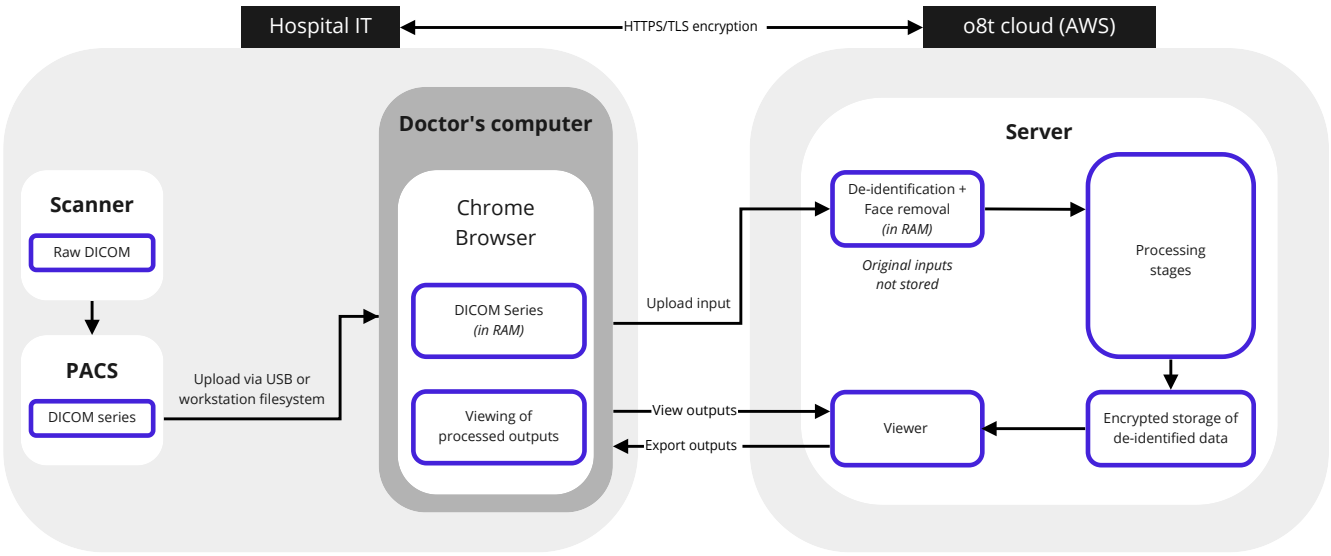


Figure 14 Data Transmission Schematic (from Chrome browser)

Data Transmission - PACS Integration Option

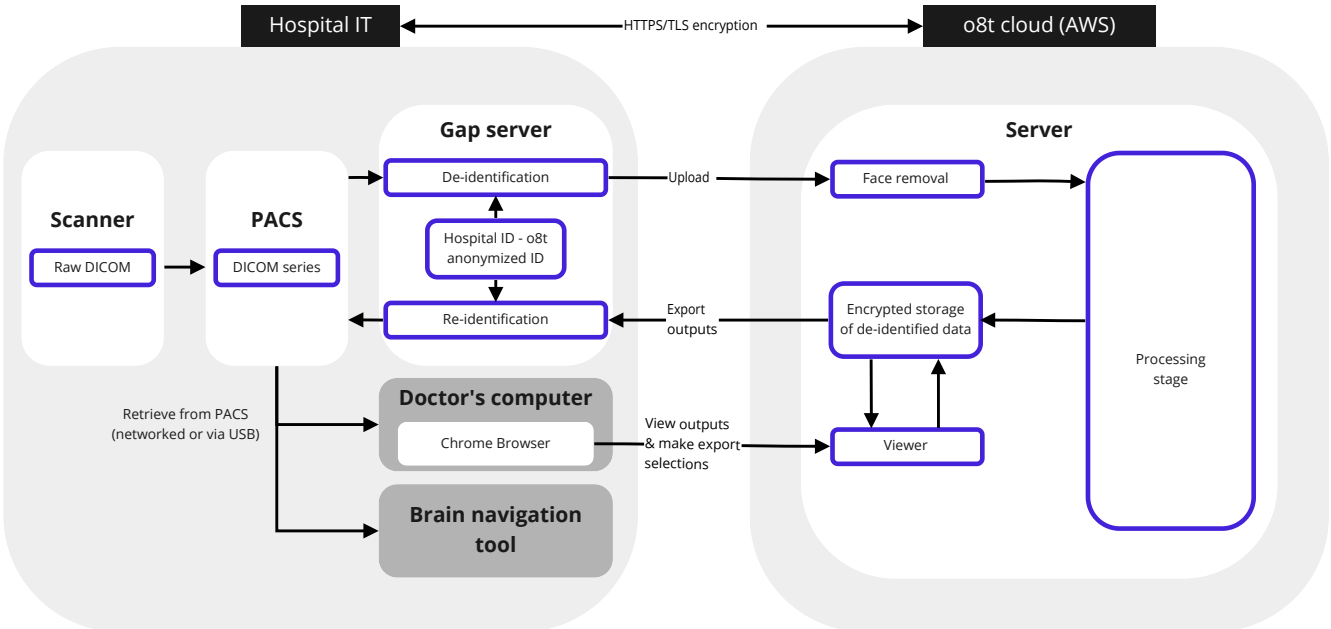


Figure 15 Data Transmission Schematic (from PACS integration)

II. Subject Deidentification

Subject deidentification is achieved in the primary datasets in the following ways:

1. Brain MRIs are deidentified through removal of the PHI fields from the DICOM header. All data being transmitted to cloud servers and gap servers are secured through end-to-end encryption.
2. Natural language processing data is deidentified using a novel machine learning algorithm which identifies and removes PHI data, specifically name, date of birth, address, referring doctor and MRN. Note MRN is used to generate the encryption code to link the NLP file to the scans and neurocognitive data. Specifically, this process involves the following steps:
    a. The process begins immediately following upload and before any data has been stored (either in the cloud or in the gap server).
    b. The first step is the conversion of raw HL7, PDF, DOC or JPG files to JSON text.
    c. The JSON file is then parsed for names from a library of 15 million first and last names from Indo-European, Anglicized Arabic, Anglicized Chinese, and Anglicized Dravidian Language names, which were created by Omniscient from several existing libraries.
    d. MRN's are identified using contextual clues such as the header MRN or patient number, numerical length, and position within the document.
    e. Address is removed by a proprietary machine learning algorithm which is trained to recognize the format of an address in most countries.
    f. The date of birth is identified by recognizing format of a date in the document.

    These algorithms were validated on a dataset of 1000 electronic medical records which were annotated by a physician, and which covered a variety of medical disciplines. The rate of identifying and removing patient names was 99.9% with the remaining 0.1% resulting from a decision to allow names which overlapped medical terms, with the goal of avoiding over aggressively cleaning medical terms from the record. For example, a name like "John Head" would be deidentified to _____Head. The rate of eliminating all other forms of PHI was 100% in this validation study.
3. Following deidentification, a series of algorithms are used to convert the remaining JSON data into a document database in MongoDB format in the cloud in a way which tags the various aspects of the data (type of document,

timing of event, section of the document the word came from, medical significance of the word) can be queried from an interface using a GUI implementation of the SQL query language.

4. To link data from the neurocognitive testing platform to existing or future scans, either a patient's MRN or a custom identifying string of characters can be used. These values are encrypted in transmission prior to de-identifcation. No other PHI is handled in this process.

III. Data Collection and Management

All research data (specifically imaging, neurocognitive testing results, and NLP data from EMR records), will be stored in the cloud platform which is a highly secured data structure inside of the Amazon Webservices (AWS) cloud environment in its country of origin.

Legal ownership and usage rights are addressed in Omniscient Terms and Conditions, but in summary the ownership of the data uploaded into the platform remains with the institution.

It is possible to delete the entire dataset for a participating site at any time, if desired, with a written request, however due to the deidentification process, it is not possible to identify and delete individual subject data from the database as there will be no method for us to identify the subject.

Records Retention

The study data will be retained at least 5 years in the AWS platform, or until a request is made to delete a sites data.

6. REFERENCES

D. Tournier, S. Mori, A. Leemans, Diffusion tensor imaging and beyond. *Magn Reson Med* 65, 1532-1556 (2011).

P. G. Batchelor, M. Moakher, D. Atkinson, F. Calamante, A. Connelly, A rigorous framework for diffusion tensor calculus. *Magn Reson Med* 53, 221-225 (2005).

D. Tournier, F. Calamante, A. Connelly, Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. *Neuroimage* 35, 1459-1472 (2007).

R. G. Briggs et al. , A Connectomic Atlas of the Human Cerebrum-Chapter 14: Tractographic Description of the Frontal Aslant Tract. *Oper Neurosurg (Hagerstown)* 15, S444-s449 (2018).

Schilling et al. , Can increased spatial resolution solve the crossing fiber problem for diffusion MRI? *NMR Biomed* 30 (2017).

Calamuneri et al. , White Matter Tissue Quantification at Low b-Values Within Constrained Spherical Deconvolution Framework. *Front Neurol* 9, 716 (2018).

D. Becker et al. , Going Beyond Diffusion Tensor Imaging Tractography in Eloquent Glioma Surgery-High-Resolution Fiber Tractography: Q-Ball or Constrained Spherical Deconvolution? *World Neurosurg* 10.1016/j.wneu.2019.10.138 (2019).

Hyde et al. , White matter organization in developmental coordination disorder: A pilot study exploring the added value of constrained spherical deconvolution. *Neuroimage Clin* 21, 101625 (2019).

P. A. Bonney et al. , A Simplified Method of Accurate Postprocessing of Diffusion Tensor Imaging for Use in Brain Tumor Resection. *Oper Neurosurg (Hagerstown)* 13, 47-59 (2017).

D. Burks et al. , A method for safely resecting anterior butterfly gliomas: the surgical anatomy of the default mode network and the relevance of its preservation. *J Neurosurg* 126, 1795-1811 (2017).

S. Cavdar, A. Esen Aydin, O. Algin, E. Aydogmus, Fiber dissection and 3-tesla diffusion tensor tractography of the superior cerebellar peduncle in the human brain: emphasize on the cerebello-hypthalamic fibers. *Brain Struct Funct* 225, 121-128 (2020).

Vassal et al. , White matter tracts involved by deep brain stimulation of the subthalamic nucleus in Parkinson's disease: a connectivity study based on preoperative diffusion tensor imaging tractography. *Br J Neurosurg* 10.1080/02688697.2019.1701630, 1-9 (2019).

Thomas et al., Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited. *Proc Natl Acad Sci U S A* 111, 16574-16579 (2014).

Dipasquale et al., Comparing resting state fMRI de-noising approaches using multi- and single-echo acquisitions. *PLoS One* 12, e0173289 (2017).

F. Glasser et al., A multi-modal parcellation of human cerebral cortex. *Nature* 536, 171-178 (2016).

Tax et al., Recursive calibration of the fiber response function for spherical deconvolution of diffusion MRI data. *NeuroImage*, 67-80 (2014).

Henderson F, Abdullah KG, Verma R, Brem S. Tractography and the connectome in neurosurgical treatment of gliomas: the premise, the progress, and the potential. *Neurosurg. Focus* 48 (2) E6, 2020

Tunç B, Ingalhalikar M, Parker D, et al. Individualized map of white matter pathways: connectivity-based paradigm for neurosurgical planning. *Neurosurgery* 2016; 79:568-577.

Parker D, Ould-Ismail AA, Wolf R, et al. Freewater estimatoR using iNtErpolated iniTialization (FERNET): Characterizing peritumoral edema using clinically feasible diffusion MRI data. *PLoS ONE*. 2020; 15(5): e0233645.