TRUST

IMPROVING HEALTH OUTCOMES THROUGH TRUSTED DATA EXCHANGE

trustplatform.sg

"Trusted Research and Real world-data Utilisation and Sharing Tech"

<National Cohorts Office Seminar>
<26 May 2023>

Jointly developed by:

MINISTRY OF HEALTH SINGAPORE

SMART NATION & DIGITAL GOVERNMENT OFFICE

IHiS Because IT's in Healthcare

GOVTECH SINGAPORE

Agenda

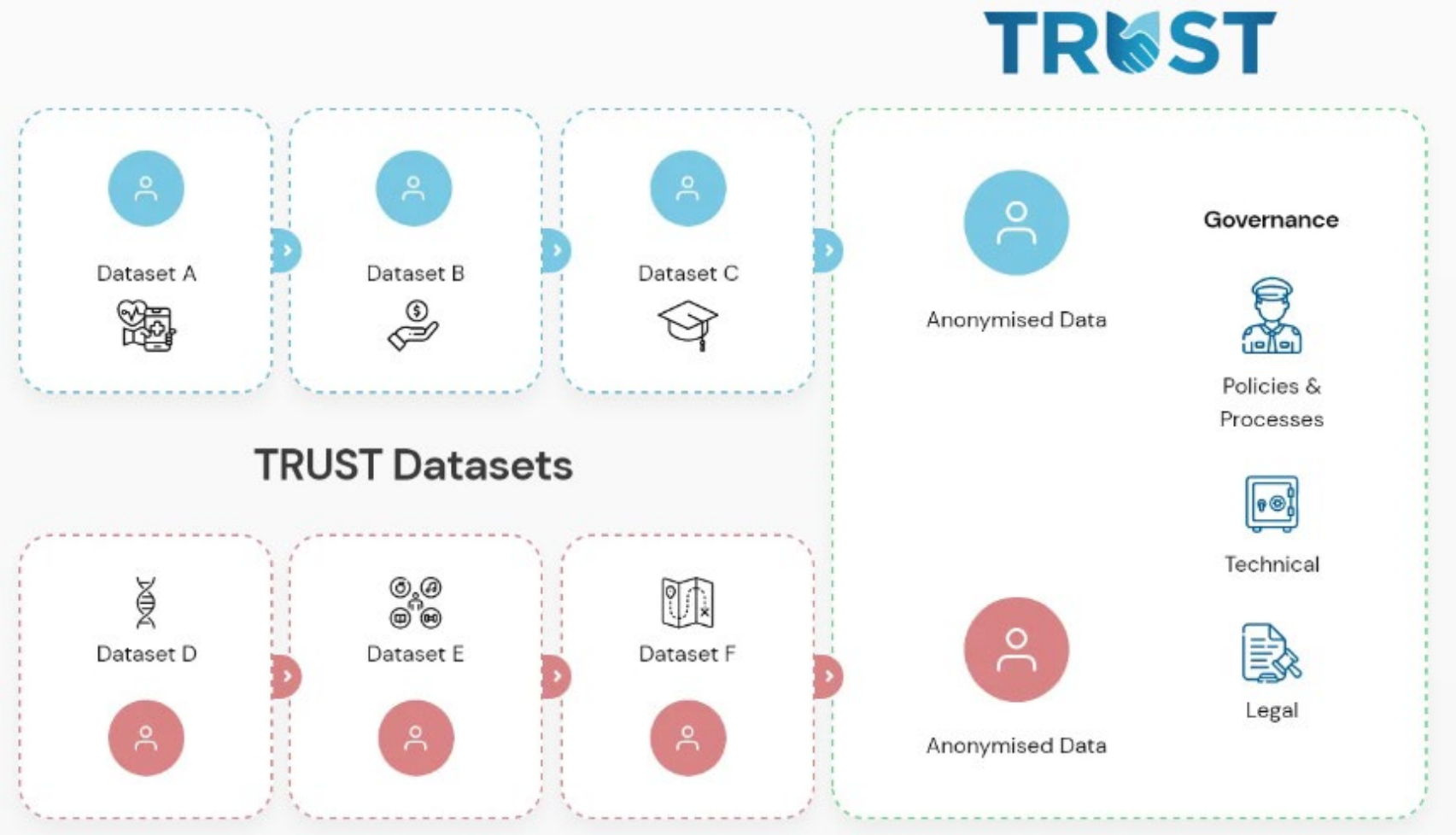
- What is TRUST
- Governance
- TRUST Framework
- How to Get Started

IMPROVING HEALTH OUTCOMES THROUGH TRUSTED DATA EXCHANGE

What is TRUST

‘Trusted Research and Real world-data Utilisation and Sharing Tech’

A National Health-related data exchange platform



Before TRUST

TIME CONSUMING AND COSTLY

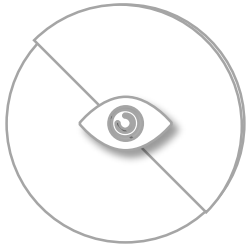
Data access could take 6 - 13 months

Extraction and transfer of data is costly

Access to datasets is time consuming due to lengthy bilateral negotiations

VARIED DATA SECURITY AND ACCESS PRACTICES

No existing platform to allow for a safe and secure platform to enable data access and linkage



VARIED DATA STANDARD, DATASETS NOT INTEROPERABLE

Varied data standards impede fusion and analytics

Data cleaning and concordance is resource-intensive

LIMITED DISCOVERABILITY

No existing platform to allow for a safe and secure platform to enable data access and linkage

With TRUST

TIME AND COST SAVINGS^

No need for bilateral negotiations with data owners

Data access could be within 6-8 weeks^ due to TRUST pre-agreements with data contributors

^estimated savings. POC phase will provide us with better sensing.

DATA GOVERNANCE AND LEGAL COMPLIANCE

Trusted Third Party (TTP) to manage de-identification

Secure environment for data fusion, access and analysis with safeguards

INTEROPERABILITY AND FLEXIBILITY

Standardized, cleaned and concorded data that **ensure interoperability and analytics**

Internationally recognised OMOP CDM standards

INNOVATION AND INSIGHTS

Access to **library of datasets** e.g. clinical, socio-economic, phenotypic data **that would otherwise not be easily accessible**

TRUST is being developed over 3 phases

Apr 2020 – Mar 2022

TRUST ("Proof-of-Concept")

- Develop pilot data sharing and fusion platform with selected datasets and use cases
- Establish "permissibility to share" and "rules of engagement": principles of rights of use of data, management of IP and associated benefits sharing
- Datasets: Genomic data, Circumscribed MOH Clinical data, Lifestyle (wearable data), Chronic disease screening data

FY 2022 to Mid-2023

TRUST ("Proof-of-Value")

- Scale up to a fully-operational technical platform
- Expand datasets to other research and administrative databases
- Expand range and depth of use cases
- Pilot access by industry

Mid-2023 onwards

TRUST ("Proof-of-Scale")

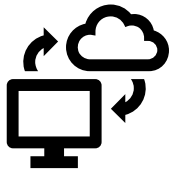
- National Health-related Data Exchange platform to facilitate data sharing and usage between the between Government, researchers and industry.
- Continual onboarding of datasets (such as deep phenotypic data, full genome sequencing data etc.) and generate use cases

Establish
scalable minimum infrastructure

Enrich
diversity of datasets

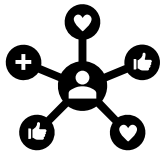
Enhance
SG's differentiating advantage

TRUST Phase I: Apr 20 – Mar 22



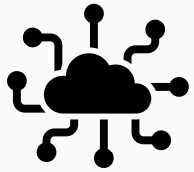
Technical

- Developed pilot data sharing and fusion platform, with data exploitative environment hosted on the Government Commercial Cloud.
- National-scale genotype-phenotype analysis could be carried out, involving research and real-world datasets across multiple institutions and consented research cohorts.



Governance

- Established governance for data access and oversight through the setting up of interim TRUST Data Access Committee (DAC).
- Developed national level pre-agreements representing a game-changing approach in terms of how publicly funded research data could be used and shared at scale in an effective and safe manner.



Data

- Worked out governance and established access to various real world datasets.
- Initiated a national level data cleaning effort, adopting OMOP to progressively clean the health related datasets as well as the clinical research datasets.

Phase I Use Cases

Use Case #1: How lifestyle choices affect the likelihood of chronic disease based on a person's genetic information (polygenic risk score)

The study enabled HPB to better understand the health outcomes of their programmes, and allow their prevention programmes to be customised to individuals of higher risk.

Linkage of:

- Polygenic Risk Score (from the National Precision Medicine Programme, SG_10K Health dataset)
- Lifestyle (wearable data) and chronic disease screening data (from HPB)
- Clinical data (from MOH)

Analysts from NPM/SG10K and HPB

Use Case #2: Determining which genetic variants are pathogenic and correlate with common inherited disorders to improve diagnosis

The study allowed for more precise diagnosis of these diseases, especially in patients of Asian ancestry as there is a general lack of data from populations of Asian ancestry in existing datasets.

Linkage of:

- Variant Calling Files (from the National Precision Medicine Programme, SG_10K Health dataset)
- Clinical data (from MOH)

Analysts from NPM/SG10K

TRUST Phase II: operational in Nov 2022

- TRUST website is live. Researchers can sign up as TRUST member to browse the data catalogue and submit data requests
- More real-world data and strategic research data are made available to support a wider range of research use cases

Real-world data

- **MOH data**
- Clinical & National Disease Registry
- **Government Administrative data (Single Source of Truth, SSOT)**
- HDB Housing / MOE Education / ICA Birth and Death / IRAS Annual Property Value
- **HPB data**
- Lifestyle (wearable data) and health screening data (e.g. HPB programmes)

Strategic research datasets

- Genomic data (e.g. PRECISE)
- Deeply characterised longitudinal population cohorts (e.g. SG100K, GUSTO)
- Well characterised disease cohorts (e.g. Singcloud, ATTRaCT)

Current technical capabilities and technologies

	FIRST PHASE Proof of Concept	SECOND PHASE Proof of Value	NEXT PHASE* Proof of Scale

Business Tools	▪ **Data manipulation scripts**
	▪ **GWAS (**HAIL format and scripts**)**
		▪ **Data Dictionary** CKAN (Trial – selected tables)
		▪ **Data Standardization** OMOP CDM
		▪ **Visualization** Python/R libraries
		▪ **OHDSI Toolset** ATLAS, DQD, USAGI, WHITERABBIT, RABBIT-IN-A-HAT (Trial – data cleaning environment)
		▪ **Genomics Platform (**Trial)
		  ▪ **Data Formats** HAIL, VCF, GVCF, GFF3, HD5
		  ▪ **Engines/Toolsets** Population genomics engine, Clinical Genomics engine
			▪ **Metadata-exchange**
			▪ ***Visualization and BI Tools***
			▪ ***Statistical tools*** *– SAS*
			▪ ***Machine learning*** *– WEKA tool*
			▪ ***Image analysis*** *– Matlab, Octave*
			▪ ***Collaboration*** *– Shareable notebooks,*
			▪ ***Automation / pipelines*** *– IHPC Modstore*
			▪ ***Machine learning*** *- Tensorflow*

Platform Tools	▪ **Data Ingestion**
	  ▪ SFTP
	▪ **Privacy Preserving Tech**
	  ▪ Anonymization script
	  ▪ Fusion script
	▪ **Data Science**
	  ▪ Jupyter notebooks
	  ▪ Sagemaker
	  ▪ AWS EMR
		▪ **Data Ingestion**
		  ▪ Remote S3
		▪ **Privacy Preserving Tech**
		  ▪ Anonymization service
		  ▪ Fusion service
		  ▪ Synthetic data (Trial)
		▪ **Data Science**
		  ▪ Jupyter notebooks
		  ▪ Sagemaker - additional libraries
		  ▪ AWS EMR
		▪ **Billing**
		  ▪ Reporting on usage
		▪ **Code Repository/ Versioning**
		  ▪ AWS Code commit
			▪ **Data Ingestion**
			  ▪ APIs
			▪ **Central and Sector architecture**
			  ▪ EnTRUST common services
			▪ **Privacy preserving Tech**
			  ▪ Synthetic data
			  ▪ Federated analysis
			▪ ***Data Science Development environment –***
			  ▪ *R studio*
			▪ **Billing**
			  ▪ Self service portal

Indicative list - subject to changes with user engagement

What are limitations of TRUST Phase II

- Current recency of available real-world datasets is 2 years (future phases will aim to shorten this to 1 year)
- TRUST does not yet support images or unstructured data (targeted implementation by FY2024)
- Federated data analysis is not yet supported (targeted trial by FY2024)

TRUST Governance

Safeguards

TRUST adopts the Five Safes Framework and deploys synergistic technical solutions across the data lifecycle (i.e. from data acquisition to destruction) to maximise data utility while ensuring secure use of data on TRUST.

SAFE PURPOSE

Before any research is allowed on TRUST, the TRUST Data Access Committee (DAC) will review the research requests.

SAFE PEOPLE

Individuals with access to TRUST must have the appropriate credentials and only work on approved research.

SAFE SETTINGS

Data on TRUST is stored in a secured environment with government-standard security measures.

SAFE DATA

Data on TRUST are accessed and used based on permission granted and are anonymised to reduce re-identification risks.

SAFE OUTPUT

Drafts of any output to be published by data requestor must be provided to TRUST first, for review and checks on any re-identification risks

TRUST DAC

Dr Cheong Wei Yang (Chair)
DS (Technology), MOH

Ms Weng Wanyi
D, GDO, SNDGO

Ms Lim Yi Ding
D, DOS/TC

A/Prof Yeo Khung Keong
Dy GCMIO (Research), SHS

A/Prof Ngiam Kee Yuan
GCTO, NUHS

Prof Benjamin Seet
Dy GCEO (Education & Research), NHG

Prof Chng Wee Joo
Vice Dean (Research), YLLSOM, NUS

Prof. Nicholas Graves
DD, HSSR, Duke-NUS

Dr Sebastian Maurer-Stroh
Executive Director, BII, A*Star

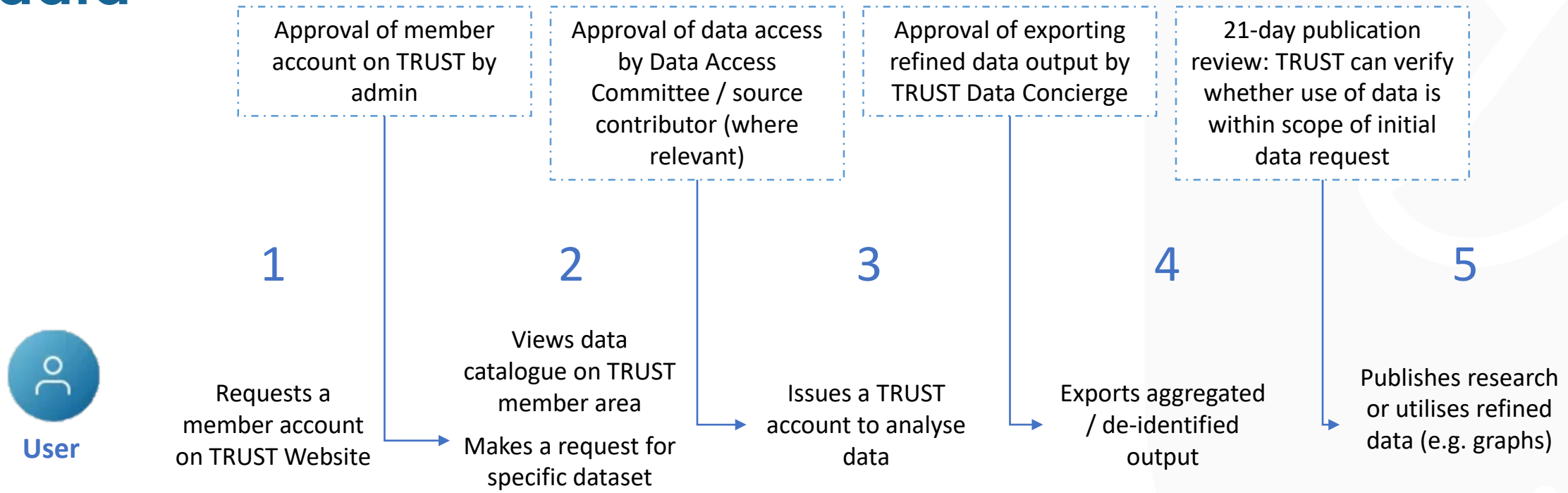
Prof John Chambers
Prof, CVD Epi, NTU
CSO, PRECISE

Prof Julian Savulescu
D, Centre for Biomedical Ethics (Ethics Domain)

Prof Simon Chesterman
Dean, NUS Law (Legal Domain)

Ms Anita Fam
President, NCSS; Chairman, Assisi Hosp

Ensuring Data User's expeditious and safe access to data



TRUST users will only be able to export approved aggregated / de-identified output (validated by TRUST Data Concierge) from TRUST.

TRUST Framework

TRUST operates on two models

Model 1

Technical platform

A safe and secure platform for collaborators to **share their own data[1] (not TRUST data)**

1. TTP for collaborators to share **their own data** with each other.
2. Linkage of **data from collaborators**
3. Exploitation environment to access and analyse the linked datasets
4. Export of insights

Model 2

The portal to rich and quality datasets

A safe and secure platform to access datasets[2] that are available through TRUST

1. TTP to access **TRUST datasets** and own research datasets
2. Linkage of multiple **TRUST datasets**, including own research datasets
3. Exploitation environment to access and analyse the linked datasets
4. Export of insights

1. Depends on researchers own research collaboration agreement to define data sharing terms.
2. TRUST datasets are available for use and access based on TRUST pre-agreed terms with data contributors. Access needs to be approved by TRUST DAC.

Ensuring Data Contributor's interest are protected through "Pre-agreements" established upfront

1

TRUST establishes pre-defined principles and data sharing governance / process with various Contributors upfront.

This will allow for TRUST DAC to approve data access on behalf of data contributors.

2

The pre-agreements cover:

- accountability of the different parties in the data sharing process
- periods of exclusive use
- consent and ethics approval
- publication and attribution
- IP and benefit sharing, etc.

3

Reduces the lengthy bilateral negotiations and speeds up data access.

Partnering TRUST as a Data Contributor

1. **Improves data quality and interoperability**
2. **Contributor retains sole use of their data during Exclusivity Period**
3. **Gain access to real world data**
4. **Expanded collaboration network after Exclusivity Period**

Partnering TRUST as a Data Contributor

1

Improves data quality and interoperability

- Receives support from TRUST's **Central Data Cleaning team** to clean their data to a **harmonised format** interoperable with international standards
- **Improves their data quality and keep their cleaned dataset**. Data Contributor retains the free and unfettered right to use the clean dataset for any purpose outside of TRUST

2

Contributor retains sole use of their data during Exclusivity Period*

- **During the Exclusivity Period***, Data Contributor retains exclusive use of own datasets and decide who can access the data

* Note: Concept of "Exclusivity Period" introduced as a NMRC data sharing term for large-sized grants e.g. LCGs since 2019, where Exclusivity period typically refers to either 24 months after the grant ends, or 12 months after publishing, whichever is earlier.
TRUST takes reference from this concept of Exclusivity Period and exact period (18-24 months) will be defined with contributors, in accordance with the date of generation of the data or from the date of commencement of the pre-agreement if the data is already generated.

Partnering TRUST as a Data Contributor

3

Gain access to real world data

- Gains access to, and fuse with, other real world data such as clinical and government data available on TRUST, which **maximises value/ utility of their datasets** (first mover advantage)

4

Expanded collaboration network after Exclusivity Period

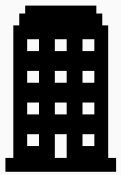
- **After Exclusivity Period,** TRUST facilitates use of datasets to a bigger group of users and this will further **expand Data Contributor's collaboration network**.
- Where Data contributor is not a collaborator, Data Contributor shall be entitled to be **granted non-exclusive and royalty-free licence to access and use the foreground IP** for its internal and non-commercial research purposes.
- Where Data Contributor is a collaborator, the ownership of such Foreground IP shall be attributed or allocated between the Data Contributor and the Data Requestor in accordance with the **terms of its own agreement.**

* Note: Concept of "Exclusivity Period" introduced as a NMRC data sharing term for large-sized grants e.g. LCGs since 2019, where Exclusivity period typically refers to either 24 months after the grant ends, or 12 months after publishing, whichever is earlier.
TRUST takes reference from this concept of Exclusivity Period and exact period (18-24 months) will be defined with contributors, in accordance with the date of generation of the data or from the date of commencement of the pre-agreement if the data is already generated.

Charging Framework (implementation by FY2024)

Framework will strike a balance between promoting widespread use of the platform and preventing abuse of use

TRUST will cover the core infrastructure development, baseline maintenance, service operation cost and the manpower to operate TRUST



Users will be required to cover the computation and storage costs for their project (i.e. cloud storage, service / tools utility etc.), e.g. AWS

How to Get Started

How to get started

For researchers from our Public Research Organisations*, you will be able to access TRUST if you are:

- An employee of an institution that has signed the *Data Request Agreement* with TRUST; and
- A bona fide researcher (verification will take place through *Pubmed ref*, *ORCID ID*, CV or institution profile page), and
- Have a verified institution email account (through email verification link)

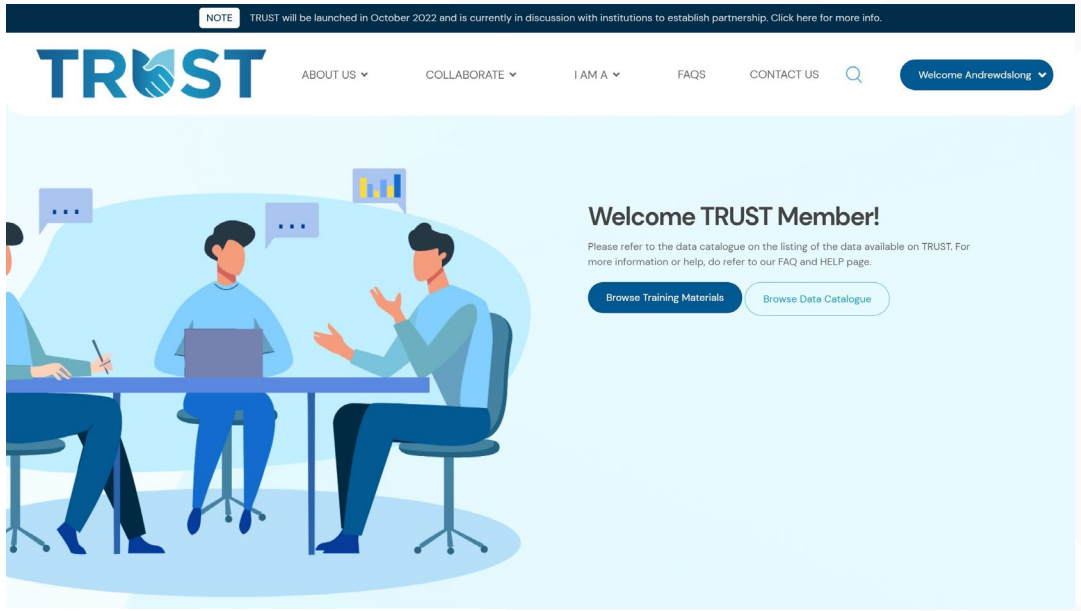
* Public sector users & researchers from Singaporean public health institutions, institutes of higher learning and publicly funded institutions

Onboarding, Training and Support

Users are supported with an onboarding programme by the TRUST team, augmented with additional resources available through the TRUST portal

- User guides
- Step-by-step video tutorials
- Community / peer forum*

*to be launched in 2024



A *TRUST Data Concierge* team supports users throughout their journey

Recommended data science competency for TRUST users

To carry out data analysis on TRUST, it is highly recommended that users should have a base level of competency in the following areas.

	Recommended Competency	Required For
Strongly Recommended	Some familiarity with Python and / or R syntax	For TRUST data research
	Some experience with handling Data Science notebook interface	For TRUST data research
	Some knowledge of how to work with Linux command line	For data discovery and transferring files from S3 bucket to notebook instance
	Some knowledge of using Cloud storage	For provisioning resources and understanding the cost
Optional	Some knowledge of using Cloud storage	For big data research analysis
	Some familiarity with Pyspark and / or SparkR syntax	For big data research analysis

TRUST Partner Institutions

In progress:

- NUHS


Thank you
Questions?