UNICORN

UNIversity data COoRdinating ceNters

Avoid Data Indigestion

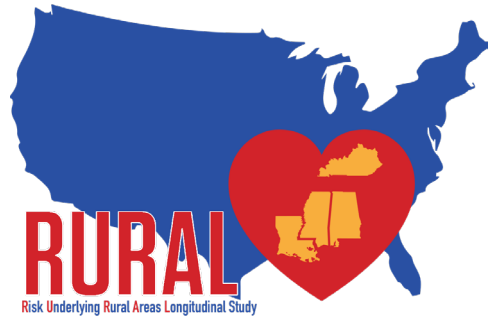
The RURAL Study

- Shawn Ballard
- Clinical Research Collaboration Unit
- Center for Clinical Epidemiology and Biostatistics
- University of Pennsylvania

crcu
Clinical Research Collaboration Unit

Penn
University of Pennsylvania

RURAL Study Overview



The Risk Underlying Rural Areas Longitudinal (RURAL) Cohort Study is a six-year **observational** research project seeking to identify why some people in the rural south may live shorter and less healthy lives.

Focusing on **10 rural counties in Alabama, Kentucky, Louisiana, and Mississippi**, the research team will examine about **4,600 residents** to study different aspects of their heart, lung, and general health.

In order to assess the overall health of study participants, several modalities of data collection are required in addition to traditional Remote Data Capture (RDC).

https://theruralstudy.org/

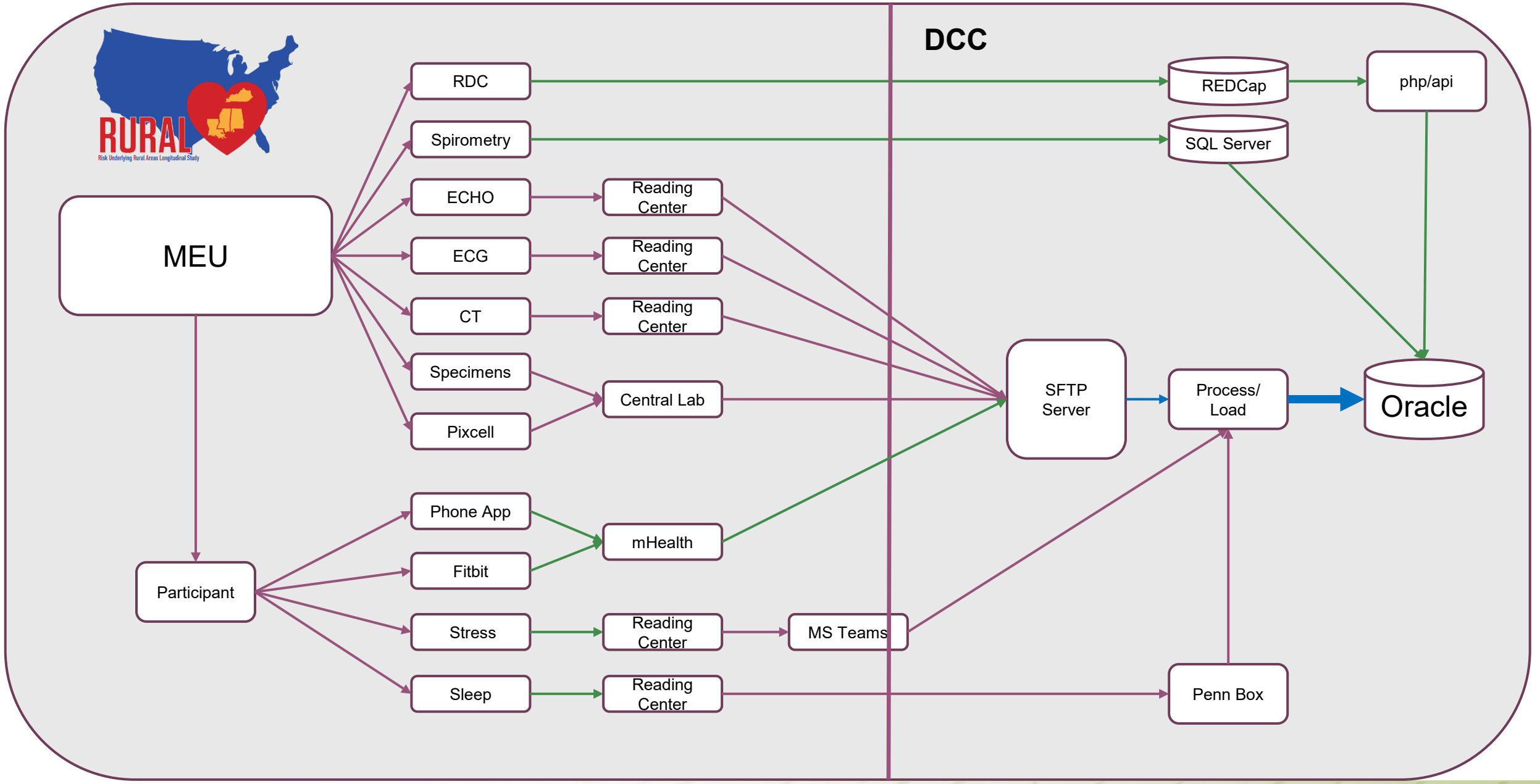
RURAL Study Overview

Participant study visits take place in the RURAL Mobile Examination Unit (MEU), a 50-foot, high-tech medical trailer, that travels to specific rural counties across Alabama, Kentucky, Louisiana, and Mississippi to reach participants in their own communities. Once the MEU arrives in a county, it remains stationary for several months at a central, accessible location

Data Ecosystem and Sources

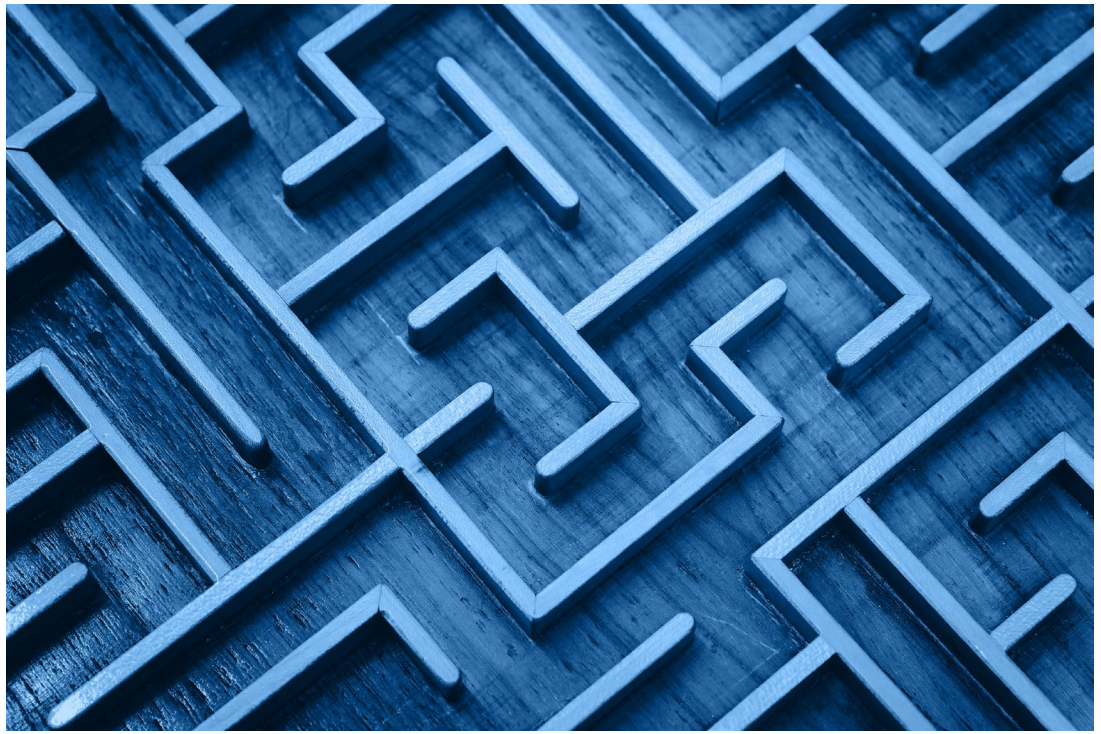
- **Data Collection on the MEU**
  - Imaging
    - ECG
    - ECHO
    - CT
  - Devices
    - Spirometry
    - Pixcell
  - RDC
    - Participant Surveys
    - CRC Data Entry
  - Specimen Collection
    - Lab Data

- **Data Collection after MEU Visit**
  - Wearables and Apps
    - Phone App (Surveys)
    - Followup (Surveys)
    - Fitbit
    - Sleep Device
    - Stress Device


RURAL
Risk Underlying Rural Areas Longitudinal Study
DCC
MEU
RDC
Spirometry
ECHO
ECG
CT
Specimens
Pixcell
Reading Center
Reading Center
Reading Center
Central Lab
Participant
Phone App
Fitbit
Stress
Sleep
mHealth
Reading Center
Reading Center
MS Teams
SFTP Server
REDCap
SQL Server
php/api
Process/ Load
Oracle
Penn Box

Key Challenges Encountered

- Expected data not received
- Identity management
- File format changes without notice
- Technical issues transferring data from MEU to reading centers
- Participant interaction with mobile app and wearables
- Issues with power to MEU
- MEU Internet Connectivity Issues



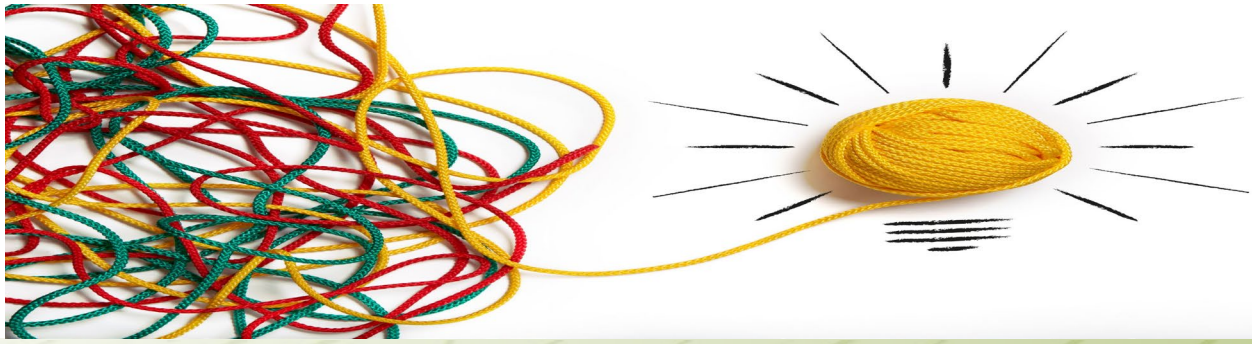
Solutions and Mitigations Implemented

- Resources added to assist participants with setting up and using phone app and mobile devices
- Regular meetings with reading centers to discuss data transfer issues
- Dashboards and QC reports created to identify missing expected data
- Real-Time QC checks to resolve identity issues
- Multiple internet options added to MEU
- Dedicated enhanced power added to MEU

Lessons learned and Takeaways

- **Pre-study visit to each site transferring data**
  - Site should present on:
    - Sample/data processing procedures
    - Identity management of samples and data
    - Data Quality procedures
  - Regular meetings to address any issues
- **Establish and document data transfer protocols**
  - Data file format (text, csv, excel, sas, xml, …)
  - File Naming Conventions
  - Data Dictionary
  - File transfer method (local ftp, remote ftp, box, email, cloud, …)
  - Incremental or cumulative data files
  - Data transfer frequency (daily, weekly, monthly, quarterly, …)
  - Contact person responsible for managing data transfers
- **Allow sites to use their established standards**
  - May be different for each site
- **Create a data flow diagram**
  - Review and get approval from all sites

- **Define and document regulatory or security requirements**
  - IRB approvals
  - DUAs
  - NIH Storage Requirements (NIST 800-171, NIST 800-53, …)
  - Mandated data repository requirements
- **Define QA/QC Plan**
  - Evaluate file per agreed upon formatting
  - Load data and integrate with central study database
  - QC checks for valid data values and identity



Northwestern Medicine®

DRPP
Dementia Risk Prediction Project

Dementia Risk Prediction Project

Denise Scholtens
John Stephen

Overall Study Goals – R61/R33

- Precision prevention approaches to identify individuals at high risk for dementia.
- Launch the Dementia Risk Prediction Project (DRPP) by pooling and rigorously harmonizing 16 prospective observational cohorts of middle and older age adults, multiple in-person assessments of clinical, genetic and behavioral risk factors, follow-up of greater than 10 years and adjudicated dementia ascertainment (R61)
- Develop and validate an accurate and personalized, dynamic dementia risk prediction model which incorporates longitudinal risk factor measurements and easily updates as new measurements are accrued (R33)

Data Ecosystem & Sources

Pre-statistical harmonization

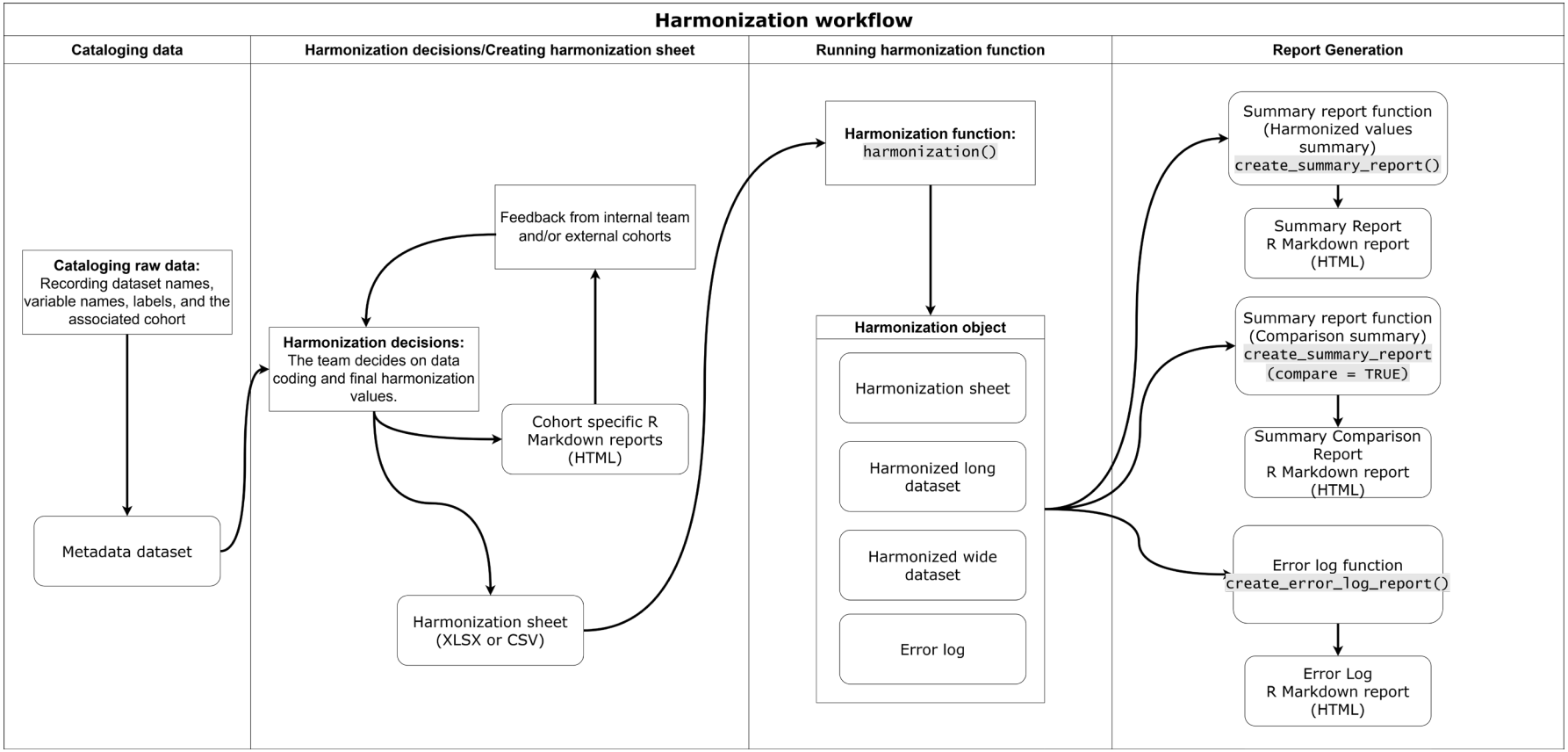
- Locally harmonize data from 14 international, longitudinal cohorts (n=100,000)

  - Age, Gene/Environment Susceptibility – Reykjavik (AGES)
  - Atherosclerosis Risk in Community (ARIC)
  - Cardiovascular Health Study (CHS)
  - Framingham Heart Study (FHS) – Original, Offspring, New Offspring Spouse, Generation 3, Omni 1, Omni 2
  - Honolulu-Asia Aging Study (HAAS)
  - Multi-Ethnic Study of Atherosclerosis (MESA)
  - Reasons for Geographic and Racial Differences in Stroke (REGARDS)
  - Sacramento Area Latino Study on Aging (SALSA)
  - Whitehall II

  - Validation Studies – 2 cohorts
    - Three-City Study
    - Rotterdam Study

Northwestern Medicine®

Harmonization workflow - psHarmonize



Northwestern Medicine®

A Harmonization instructions for height

	study	visit	item	source_dataset	source_item	code1	code_type	coding_notes	id_var	possible_ran	domain	subdomain	final_units	source_units	label	notes	source_labels
2	AGES	1	Height	ages733_drrp_first_visit_additional	HEIGHT				ID		Clinical Risk F	Height	cm	n/a			
3	ARIC	1	Height	VISIT_1	ANTA01			No change needed	ID		Clinical Risk F	Height	cm	cm			STANDING HEIGHT TO NEAR
4	ARIC	3	Height	VISIT_3	ANTC1			No change needed	ID		Clinical Risk F	Height	cm	cm			STANDING HEIGHT TO NEAR
5	ARIC	4	Height	VISIT_4	ANTD1			No change needed	ID		Clinical Risk F	Height	cm	cm			STANDING HEIGHT TO NEAR
6	ARIC	5	Height	VISIT_5	ANT3			No change needed	ID		Clinical Risk F	Height	cm	cm			Standing height (cm)
7	ARIC	6	Height	VISIT_6	ANT3			No change needed	ID		Clinical Risk F	Height	cm	cm			Standing Height (cm)
8	ARIC	7	Height	VISIT_7	ANT3			No change needed	ID		Clinical Risk F	Height	cm	cm			Standing Height (cm)
9	CHS	2	Height	CHS_main	stht_y2			Imputed 2 height values on visit 18 per CHS instruct	idno		Clinical Risk F	Height	cm	cm			STANDING HEIGHT - CM
10	CHS	5	Height	CHS_main	stht_y5			Imputed 2 height values on visit 18 per CHS instruct	idno		Clinical Risk F	Height	cm	cm			STANDING HEIGHT - CM
11	CHS	9	Height	CHS_main	stht_y9			Imputed 2 height values on visit 18 per CHS instruct	idno		Clinical Risk F	Height	cm	cm			STANDING HEIGHT - CM
12	CHS																
13	FHS - Cohort																
14	FHS - Cohort																
15	FHS - Cohort																
16	FHS - Cohort																
17	FHS - Cohort																
18	FHS - Cohort																
19	FHS - Cohort																

B Harm

	study
2	AGES
3	ARIC
4	CHS
5	FHS - Cohort 0
6	FHS - Cohort 1
7	FHS - Cohort 2
8	FHS - Cohort 3
9	FHS - Cohort 7
10	FHS - Cohort 72
11	HAAS
12	MESA
13	REGARDS
14	SALSA
15	WHITE

C Harm

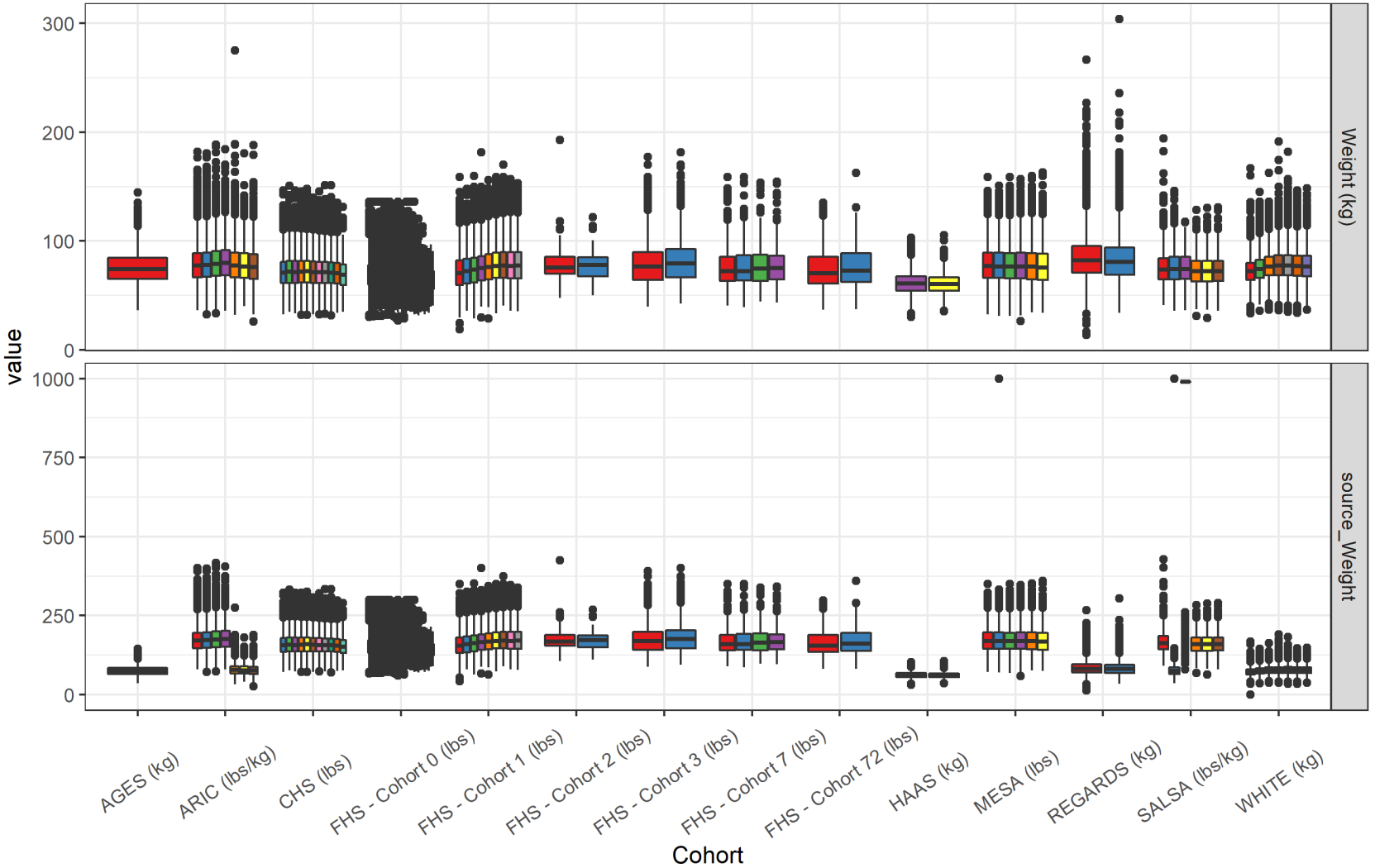
	study
1	ARIC
2	SALSA

A Harmonization instructions for height

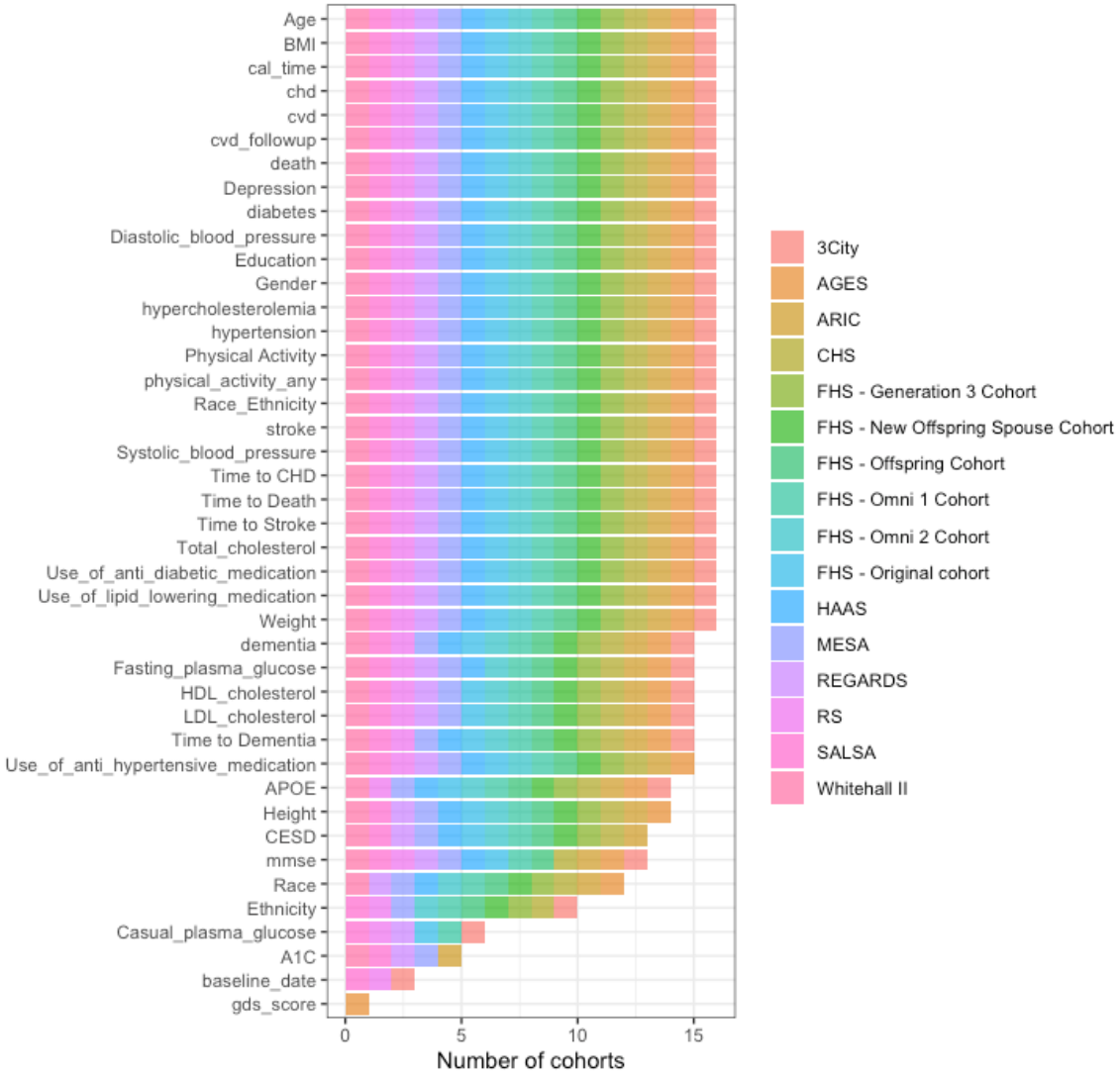
	study	visit	item	source_dataset	source_item	code1	code_type	co
2	AGES	1	Height	ages733_drrp_first_visit_additional	HEIGHT			
3	ARIC	1	Height	VISIT_1	ANTA01			No
4	ARIC	3	Height	VISIT_3	ANTC1			No
5	ARIC	4	Height	VISIT_4	ANTD1			No
6	ARIC	5	Height	VISIT_5	ANT3			No
7	ARIC	6	Height	VISIT_6	ANT3			No
8	ARIC	7	Height	VISIT_7	ANT3			No
9	CHS	2	Height	CHS_main	stht_y2			Im
10	CHS	5	Height	CHS_main	stht_y5			Im
11	CHS	9	Height	CHS_main	stht_y9			Im
12	CHS	18	Height	chs_height_v18	stht_y18			Im
13	FHS - Cohort	1	Height	vr_wkthru_ex32_0_0997s_19	HGT1	x * 2.54	function	Co
14	FHS - Cohort	4	Height	vr_wkthru_ex32_0_0997s_19	HGT4	x * 2.54	function	Co
15	FHS - Cohort	5	Height	vr_wkthru_ex32_0_0997s_19	HGT5	x * 2.54	function	Co
16	FHS - Cohort	10	Height	vr_wkthru_ex32_0_0997s_19	HGT10	x * 2.54	function	Co
17	FHS - Cohort	11	Height	vr_wkthru_ex32_0_0997s_19	HGT11	x * 2.54	function	Co
18	FHS - Cohort	12	Height	vr_wkthru_ex32_0_0997s_19	HGT12	x * 2.54	function	Co
19	FHS - Cohort	13	Height	vr_wkthru_ex32_0_0997s_19	HGT13	x * 2.54	function	Co

B Harmonization instructions for education

Harmonization results (cont.)



Harmonization results



Harmonization performed internally for 14 cohorts.

'Harmonization sheet' and the R package was provided to the 2 validation cohorts so that their harmonized data sets met our variable definitions.

DRPP products

- psHarmonize pipeline
  - Stephen et al. (2024) *psHarmonize: Facilitating reproducible large-scale pre-statistical data harmonization and documentation in R*. Patterns 5(8):101003.
- psHarmonize R package
  - https://github.com/NUDACC/psHarmonize
  - https://cran.r-project.org/web/packages/psHarmonize/index.html
- DRPP tableau dashboard
- Cloud based analysis platform
  - Outside researchers can apply to analyze harmonized DRPP data in the Apporto platform

Take home points

- Key challenges
  - Meticulous documentation of data sources and source variable definitions
  - Harmonization decision-making
  - Could rely on Common Data Elements
- What worked well
  - Harmonization sheet and R package
  - All decisions and mappings are in one place
  - More human-readable than code (?)
- Next time
  - There was a next time….
  - And another next time…
  - I would recommend psHarmonize

Northwestern Medicine®

Northwestern Medicine®

Thanks!

Lessons Learned from the Apple Heart Study

Haley Hedlin, PhD
Quantitative Sciences Unit

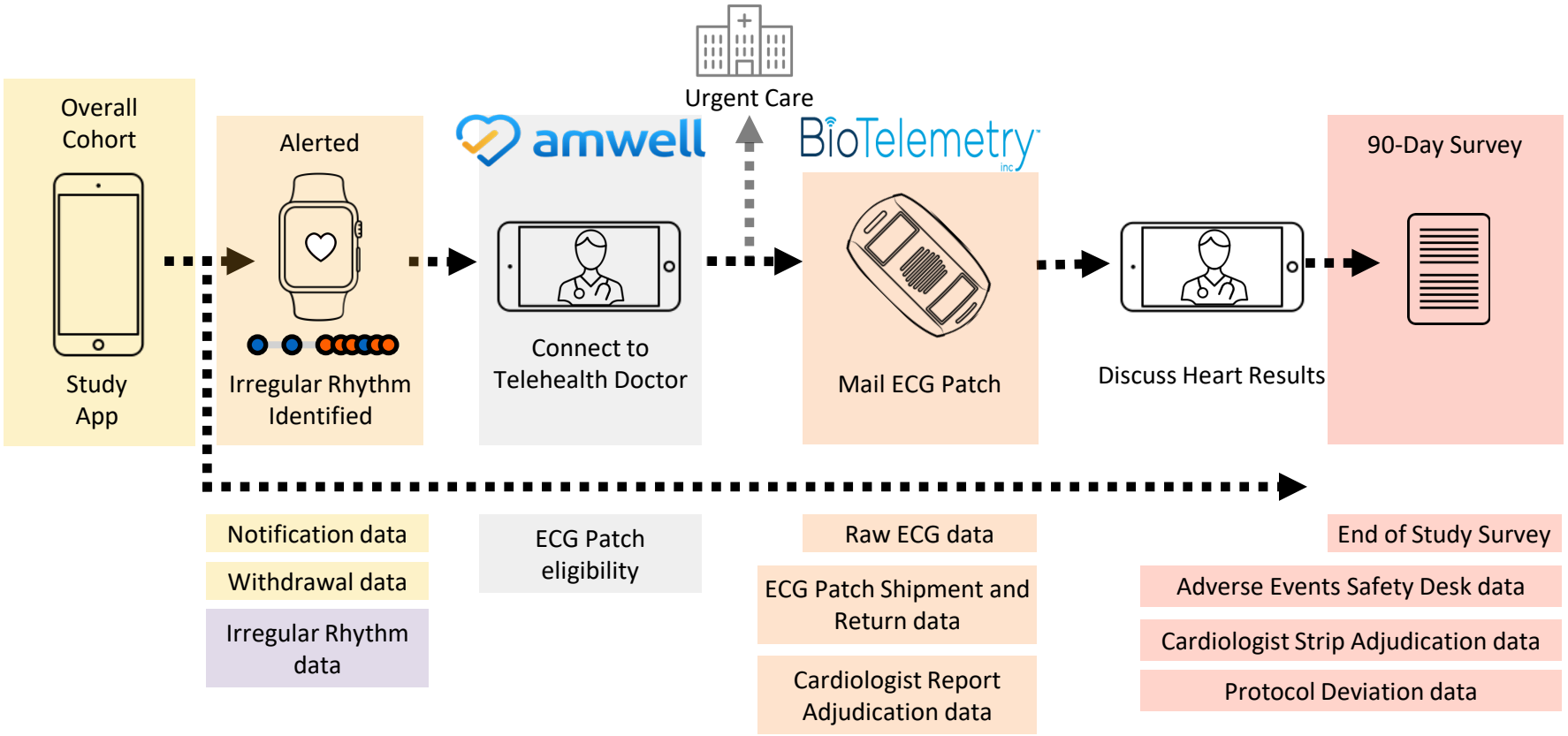
STANFORD MEDICINE

Disclosures

- My institution received funds from Apple Inc. to partially support my salary
- I receive honoraria for serving on DSMBs

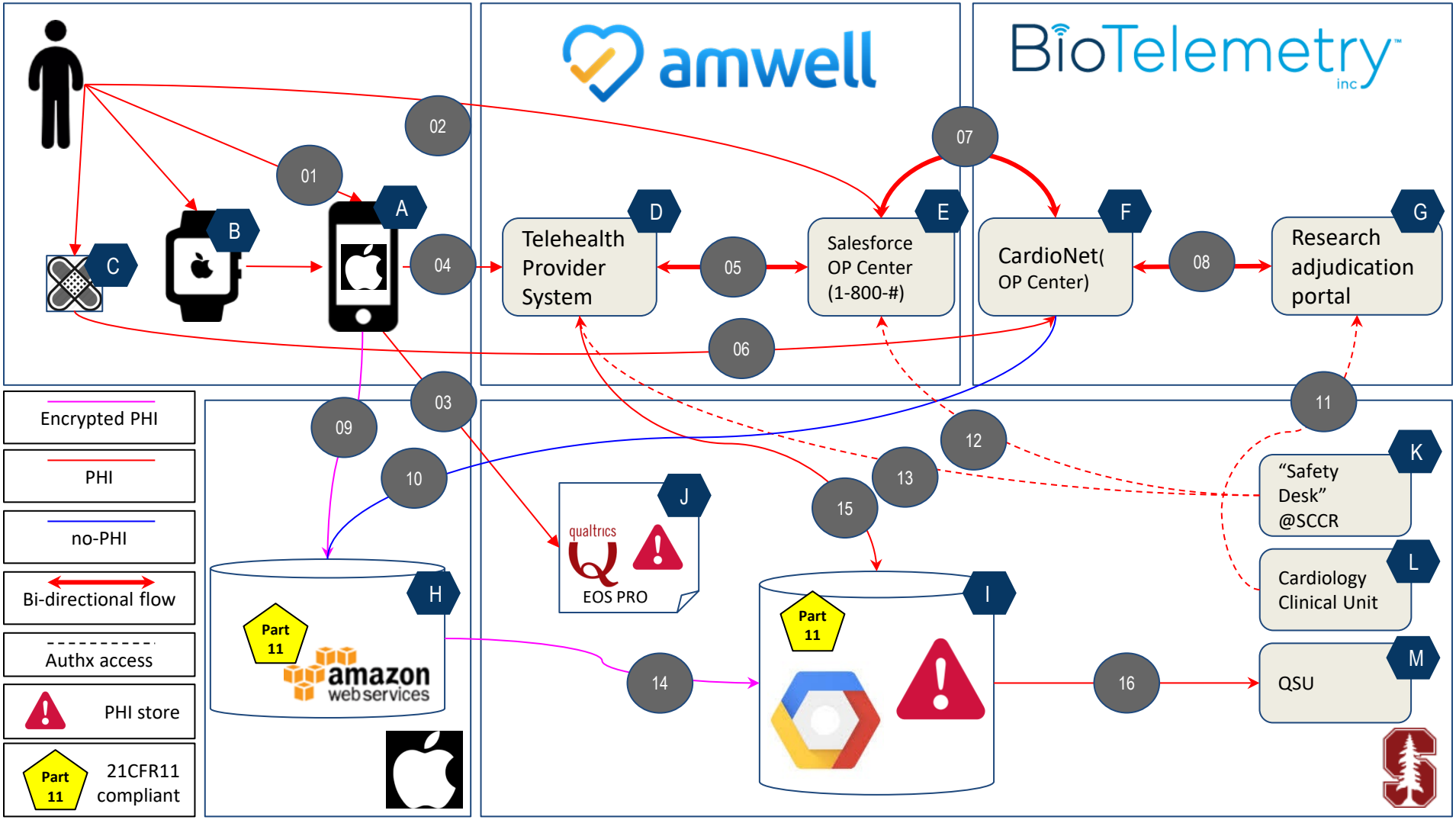
Apple Heart Study

- Decentralized, pragmatic clinical trial to evaluate the ability of the irregular pulse notification algorithm to identify atrial fibrillation (AF)

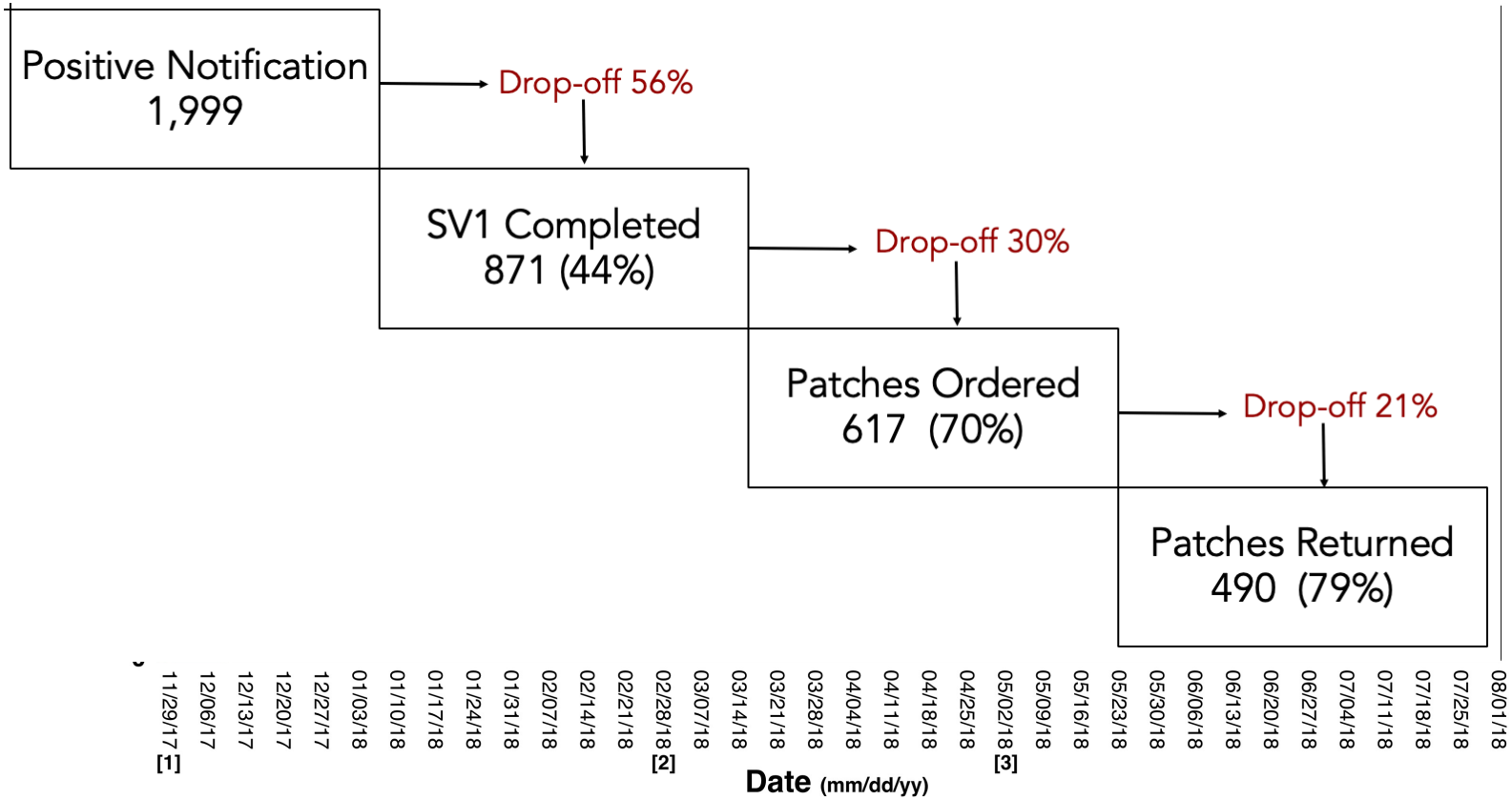


Perez MV, Mahaffey KW, Hedlin H, Rumsfeld JS, Garcia A, Ferris T, ... & Turakhia MP (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine*, *381*(20), 1909-1917.

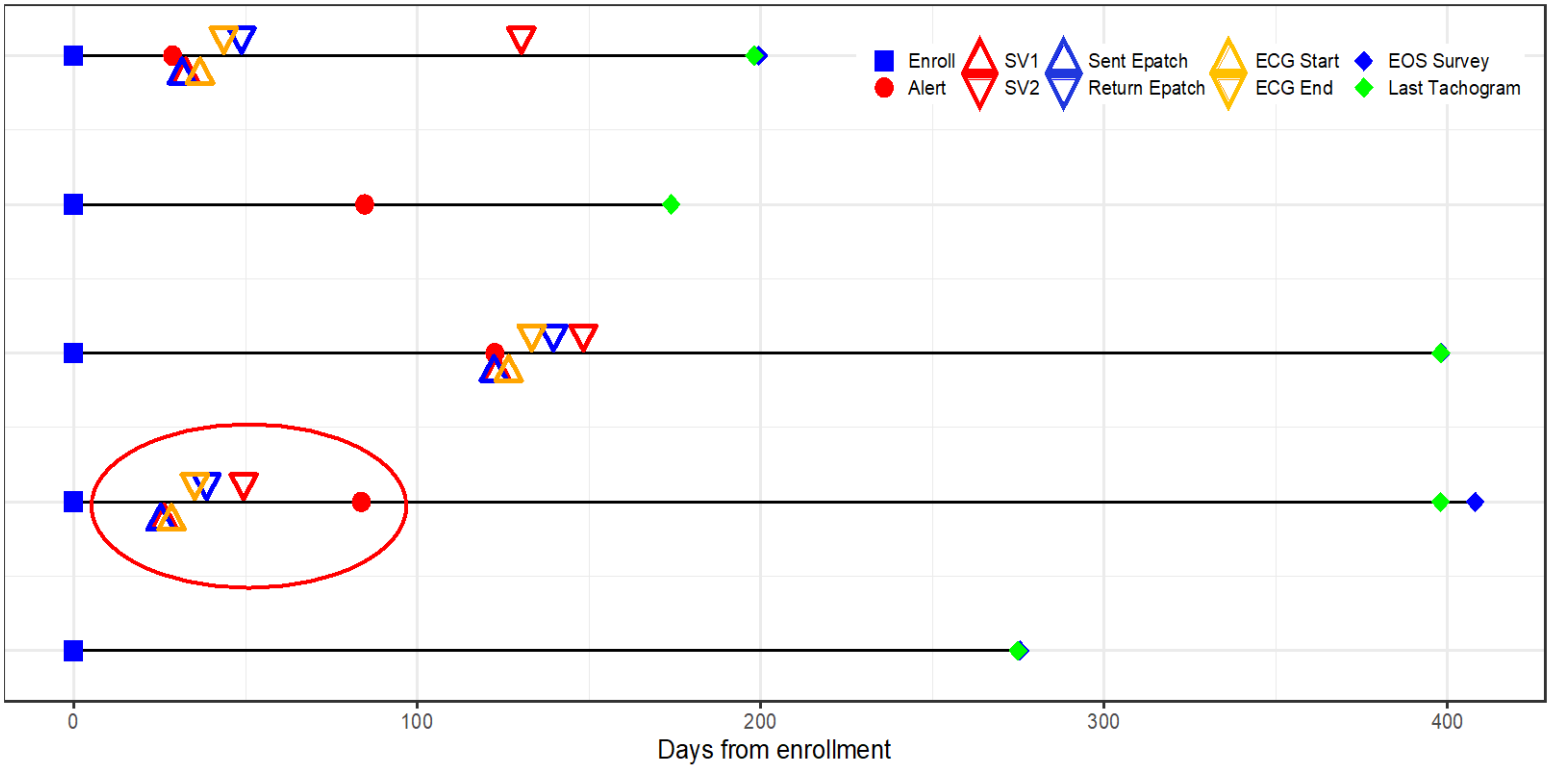
Data Pipeline



Enrollment and Monitoring



Inconsistent Temporal Ordering



Garcia A, Balasubramanian V, Lee J, Gardner R, Gummidipundi S, ... Hedlin H & Desai M (2022). Lessons learned in the Apple Heart Study and implications for the data management of future digital clinical trials. *Journal of biopharmaceutical statistics*, *32*(3), 496-510.

Participant Linkage

Deleting & reinstalling app

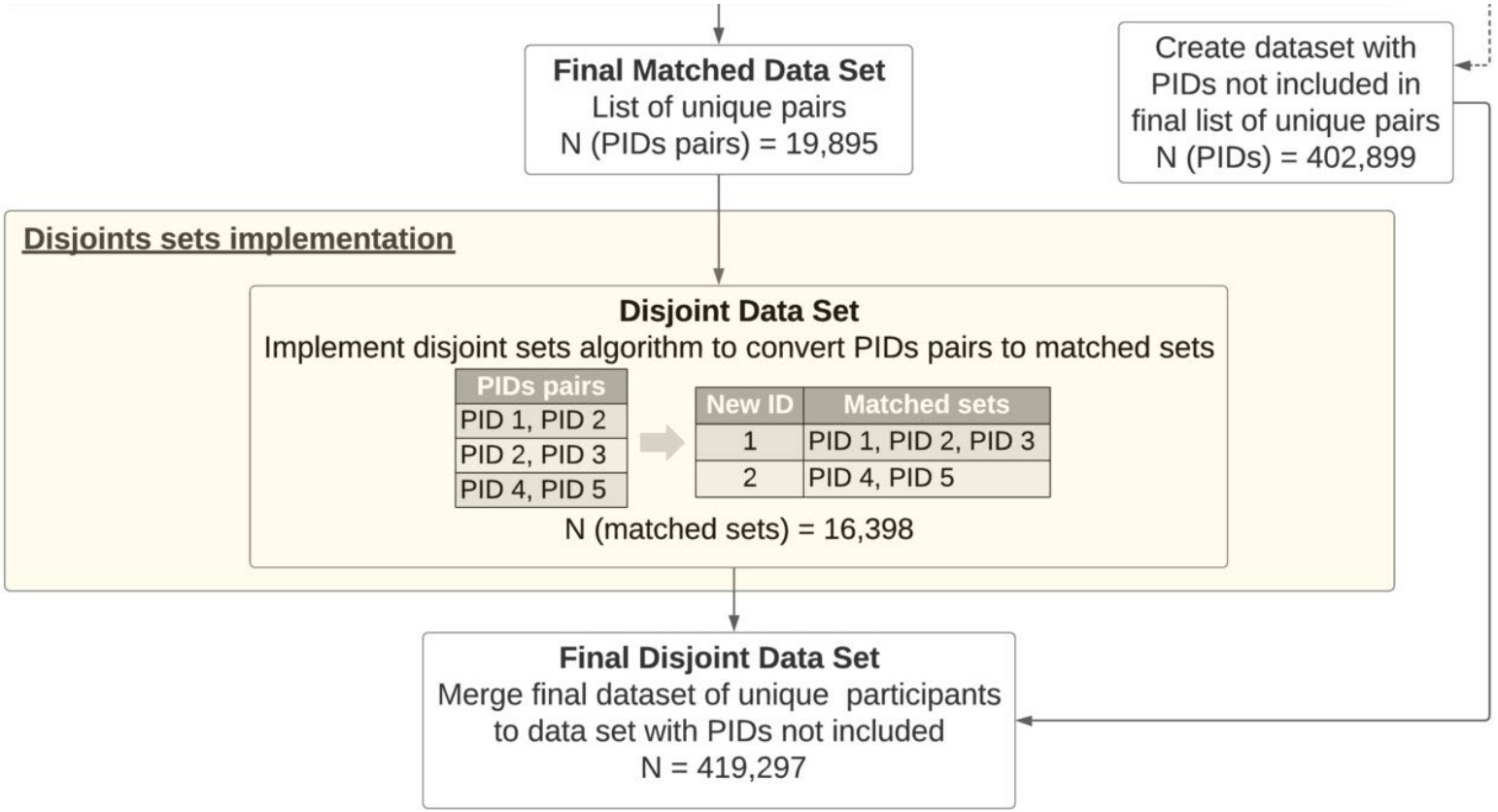
Changing devices

Device/ software issues

We developed and evaluated an algorithm to deduplicate records in real time

Garcia A, Lee J, Balasubramanian V, Gardner R, Gummidipundi S, ... Hedlin H & Desai M (2022). The development of a mobile app-focused deduplication strategy for the Apple Heart Study that informs recommendations for future digital trials. *Stat*, *11*(1), e470.

Participant Linkage



? How do we estimate and report the uncertainty of our sample size?

Garcia A, Lee J, Balasubramanian V, Gardner R, Gummidipundi S, ... Hedlin H & Desai M (2022). The development of a mobile app-focused deduplication strategy for the Apple Heart Study that informs recommendations for future digital trials. *Stat*, *11*(1), e470.

Data and Safety Monitoring

Topics of interest
Enrollment and retention numbers and trial characteristics
Adherence and loss to follow-up
Data pipeline issues
Data linkage using time stamping across platforms
Uptake or engagement of the intervention
App glitches and software updates
Data sharing and management plans
Variable and endpoint definitions from real-world data
Fairness of algorithms integrated into interventions

Bunning, B. J., Hedlin, H., Chen, J. H., Ciolino, J. D., Ferstad, J. O., Fox, E., ... & Desai, M. (2023). The evolving role of data & safety monitoring boards for real-world clinical trials. *Journal of Clinical and Translational Science*, *7*(1), e179.

Takeaways

- Perform a pilot study!
  - End-to-end testing did not identify these issues
- Timestamps on each piece of data
  - Related: Calibration and battery life
- Include a DSMB member with informatics expertise

Questions?

Haley Hedlin

hedlin@stanford.edu

med.stanford.edu/qsu

Secure Transfer of Device Data: The Sleep for Stroke Management and Recovery Trial (Sleep SMART)

CTN: NCT03812653

C Arnaud[1] , V Durkalski-Mauldin[1]

[1] The Data Coordination Unit (DCU), Department of Public Health Sciences

Medical University of South Carolina, Charleston, SC.

SCT
SOCIETY FOR CLINICAL TRIALS
47th Annual Meeting (Phoenix, AZ 2026)

Disclosures

- Sleep SMART is funded through a grant by NIH-NINDS
  - Project PIs: Drs. Devin Brown and Ronald Chervin (University of Michigan)
  - StrokeNet National Data Management Center (NDMC at MUSC)
  - StrokeNet National Coordinating Center (NCC at University of Cincinnati)
- Nox Health worked with the study team during the trial and partnered on the device data transfer to the NDMC
- Authors have nothing to disclose

Contents:

1. Study Overview
2. Data collection and frequency
3. Data transfer tools SFTP
4. Challenges Key Takeaways
5. Q & A

Study Overview

- Investigator-initiated, phase 3 multicenter, prospective randomized open-, blinded-endpoint (PROBE) controlled trial to test whether treatment of obstructive sleep apnea (OSA) with continuous positive airway pressure (CPAP) is effective for secondary prevention and recovery after stroke.

- The primary goals of Sleep SMART are to determine whether treatment of obstructive sleep apnea (OSA) with positive airway pressure starting shortly after acute ischemic stroke or high risk TIA:

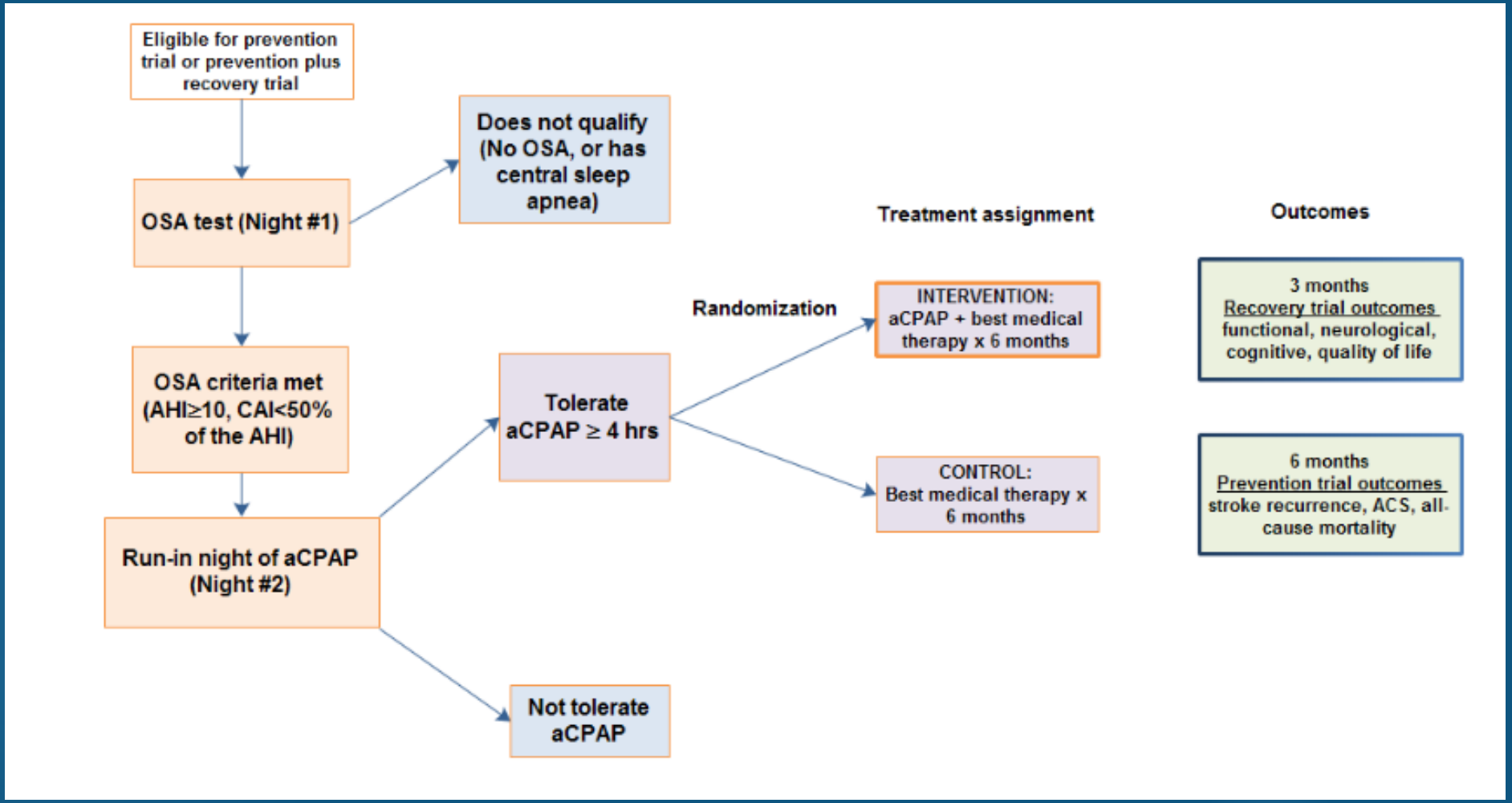
  (1) reduces recurrent stroke, acute coronary syndrome, and all-cause mortality during 6 months after the event (prevention outcome)

  (2) improves stroke outcomes at 3 months in patients who experienced an ischemic stroke (recovery outcome)

Study Overview

- Intervention Arm: 6 months of CPAP with Usual Care.
- Sites ~ 146
- Planned participants that will conduct a screening run-in night (sleep apnea test) ~15,000
- Targeted sample size – 3,062
- Device data collection
  - Sleep apnea test data to examine eligibility (Nox T3)
    - Initial test to screen for obstructive sleep apnea and rule out central sleep apnea. (centralized testing site)
  - aCPAP Data (Daily data)
    - Data collected and managed via aCPAP device (device per participant for at home nightly use for this trial)
  - aCPAP Mask Event
    - Data collected on device removal (mask on/off) during nightly use

Sleep SMART flow



Nox T3 testing at site

- a nasal cannula (nasal pressure transducer and snore sensor)
- thoracic and abdominal respiratory effort (RIP) belts
- EKG leads and finger probe (wireless pulse oximetry including pulse rate monitor).
- the chest and wrist units provide a body position sensor, microphone, and actigraph that monitors movement.

❑ Signals recorded by device uploaded to central location through a USB cable
❑ Signals centrally scored
❑ Scores sent back to the site via the device's central system
❑ Signals sent to the NDMC for analysis and storage (later to be shared publicly)

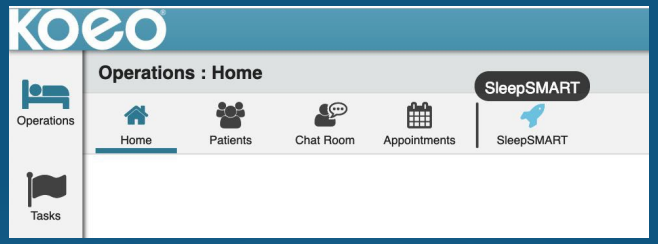
aCPAP device

- Plugs into an electrical outlet, has a humidifier, and a tube that carries air under pressure to the mask.
- CPAP device with built-in modem for remote monitoring of usage, mask leak, and other treatment related information.
- Host company uses a platform-based cloud application to provide passive remote collection of aCPAP data via cellular networks across the US.

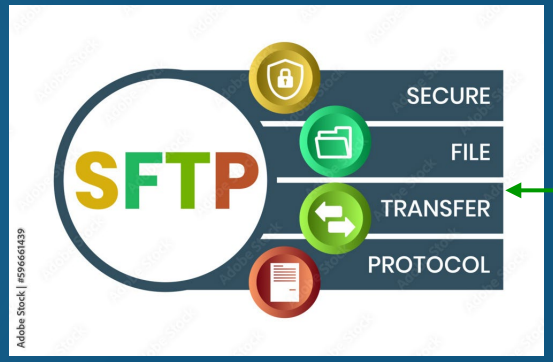
CPAP Device Data

Nightly use by participant at home for up to 6months

Secure device registration at site using assigned unique Participant ID



Data securely transferred to Device Company's portal via cellular network



The Data Coordination Unit Study Database

Data transfer tools SFTP

- SFTP (Secure File Transfer Protocol)
- What it does:
  - Securely transfers files between systems using encryption (SSH)
    - End-to-end encryption (SSH) protects data in transit with strong authentication
- How it works
  - Device / Central Site → 🔐 SFTP → DCC Central Server → Processing ➡️ Analysis team
- Why it matters:
  - Protects sensitive clinical data
  - Enables automated data transfer
  - Supports centralized data collection
  - Meets compliance requirements (HIPAA, GCP)

Data transfer tools SFTP

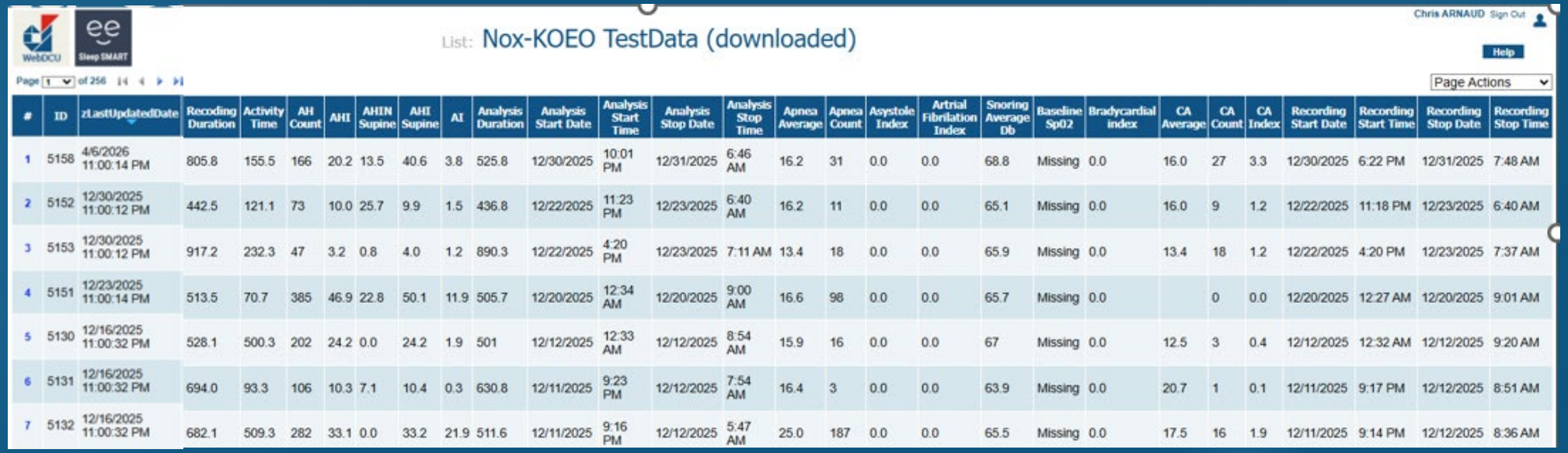
- SFTP Data Processing Workflow:
  - Automated scripts (PowerShell) securely connect to remote servers using authenticated credentials and server fingerprint validation
  - Files are downloaded over an encrypted (SFTP) connection
  - The receiving system stores incoming data and archives source files, on both receiving system and server
  - Data is loaded into MS SQL Server and integrated with existing datasets, initial data checks are performed, and notifications are sent as needed.
  - Processed data are made available for NDMC centralized monitoring and operational review.
  - Data are exported to SAS datasets for reporting and statistical analysis.
    - Device data are included in data freezes.

Data transfer tools SFTP

- Reliability & Integrity
  - Built-in data integrity checks (ensures files are not corrupted)
  - Supports resume/retry for failed transfers
  - Stable for large file transfers (e.g., device data)
- Automation Friendly
  - Easily integrated with scripts (PowerShell, batch, schedulers)
  - Enables fully automated data pipelines
  - No manual intervention required once configured and scheduled.
- Optional Comparison Line (if needed)
  - FTP: Not secure ❌
  - Email/Manual: Not scalable – can be secured but this also add another level of complexity on both central data provided and centralized data management team
  - ❌APIs: Powerful but require more development from both sides
  - ⚙️SFTP: Secure, simple, and reliable ✅

Data collection and frequency – NOX T3 data

- Sleep apnea test data (Nox T3 data)
  - Initial testing of eligibility
  - One test per participant with potential rerun/correction
  - 142 data elements collected, single file per participant.
  - Potential in this study to have 15K tests collected



WebDCU | Sleep SMART

List: Nox-KOEO TestData (downloaded)

Chris ARNAUD Sign Out | Help

Page 1 of 256 | Page Actions

#	ID	zLastUpdatedDate	Recoding Duration	Activity Time	AH Count	AHI	AHIN Supine	AHI Supine	AI	Analysis Duration	Analysis Start Date	Analysis Start Time	Analysis Stop Date	Analysis Stop Time	Apnea Average	Apnea Count	Asystole Index	Artrial Fibrilation Index	Snoring Average Db	Baseline Sp02	Bradycardial index	CA Average	CA Count	CA Index	Recording Start Date	Recording Start Time	Recording Stop Date	Recording Stop Time
1	5158	4/6/2026 11:00:14 PM	805.8	155.5	166	20.2	13.5	40.6	3.8	525.8	12/30/2025	10:01 PM	12/31/2025	6:46 AM	16.2	31	0.0	0.0	68.8	Missing	0.0	16.0	27	3.3	12/30/2025	6:22 PM	12/31/2025	7:48 AM
2	5152	12/30/2025 11:00:12 PM	442.5	121.1	73	10.0	25.7	9.9	1.5	436.8	12/22/2025	11:23 PM	12/23/2025	6:40 AM	16.2	11	0.0	0.0	65.1	Missing	0.0	16.0	9	1.2	12/22/2025	11:18 PM	12/23/2025	6:40 AM
3	5153	12/30/2025 11:00:12 PM	917.2	232.3	47	3.2	0.8	4.0	1.2	890.3	12/22/2025	4:20 PM	12/23/2025	7:11 AM	13.4	18	0.0	0.0	65.9	Missing	0.0	13.4	18	1.2	12/22/2025	4:20 PM	12/23/2025	7:37 AM
4	5151	12/23/2025 11:00:14 PM	513.5	70.7	385	46.9	22.8	50.1	11.9	505.7	12/20/2025	12:34 AM	12/20/2025	9:00 AM	16.6	98	0.0	0.0	65.7	Missing	0.0		0	0.0	12/20/2025	12:27 AM	12/20/2025	9:01 AM
5	5130	12/16/2025 11:00:32 PM	528.1	500.3	202	24.2	0.0	24.2	1.9	501	12/12/2025	12:33 AM	12/12/2025	8:54 AM	15.9	16	0.0	0.0	67	Missing	0.0	12.5	3	0.4	12/12/2025	12:32 AM	12/12/2025	9:20 AM
6	5131	12/16/2025 11:00:32 PM	694.0	93.3	106	10.3	7.1	10.4	0.3	630.8	12/11/2025	9:23 PM	12/12/2025	7:54 AM	16.4	3	0.0	0.0	63.9	Missing	0.0	20.7	1	0.1	12/11/2025	9:17 PM	12/12/2025	8:51 AM
7	5132	12/16/2025 11:00:32 PM	682.1	509.3	282	33.1	0.0	33.2	21.9	511.6	12/11/2025	9:16 PM	12/12/2025	5:47 AM	25.0	187	0.0	0.0	65.5	Missing	0.0	17.5	16	1.9	12/11/2025	9:14 PM	12/12/2025	8:36 AM

Data collection and frequency – aCPAP data

- aCPAP Data (daily data)
  - Delivery of a daily consolidated file containing 23 data elements for each participant (actively tracked).
    - Current day data with rolling historical data (e.g., prior 4 days), including corrections to previously submitted data
  - Over 243K records created and updated throughout project.

Chris ARNAUD Sign Out

List: Nox-KOEO PAPData (downloaded)

Help

of 366

Page Actions

Usage Minutes	Device Name	Device Mode	Min Auto Set Pressure Setting	Max Auto Set Pressure Setting	AHI	AI	HI	AI	CI	OI	UI	Mask Leak Median	Mask Leak 95th Percentile	Mask Leak Max	Pressure Median	Pressure 95th Percentile	Pressure Max	EPR type	EPR Level	Autoset Response	Key	Last Updated Date	Inserted Date
282	AirSense 10 AutoSet	AUTOSET	8.8	20	0.4	0	0.4	0	0	0	0	0.1	0.2	0.42	11.16	13.08	14.64	OFF	ONE	OFF	100004/2020-08-09		11/17/2022 12:35:52 PM
450	AirSense 10 AutoSet	AUTOSET	8.8	20	0.9	0	0.9	0	0	0	0	0.3	0.58	0.7	11.64	15.72	16.68	OFF	ONE	OFF	100004/2020-08-08		11/17/2022 12:35:52 PM
527	AirSense 10 AutoSet	AUTOSET	8.8	20	0.3	0	0.3	0	0	0	0	0.34	0.62	0.94	11.64	16.8	18.36	OFF	ONE	OFF	100004/2020-08-07		11/17/2022 12:35:52 PM
0	AirSense 10 AutoSet	AUTOSET	8.8	20														OFF	ONE	OFF	100004/2020-08-06		11/17/2022 12:35:52 PM
0	AirSense 10 AutoSet	AUTOSET	8.8	20														OFF	ONE	OFF	100004/2020-08-05		11/17/2022 12:35:52 PM
0	AirSense 10 AutoSet	AUTOSET	8.8	20														OFF	ONE	OFF	100004/2020-08-04		11/17/2022 12:35:52 PM
0	AirSense 10 AutoSet	AUTOSET	8.8	20														OFF	ONE	OFF	100004/2020-08-03		11/17/2022 12:35:52 PM
[illegible]	AirSense 10	[illegible]																			100004/2020-08-		11/17/2022

Data collection and frequency – Mask Data

- Daily data with multiple mask events per day with potential corrections.
  - Current day data with rolling historical data (e.g., prior 4 days), including corrections to previously submitted data
- Over 690K records created and updated throughout project .

List: Nox-KOEO MaskEvent (downloaded)

Chris ARNAUD Sign Out

Help

Page Actions

Data Date	Event Type	Event Time	Last Updated Data	Inserted Date
2019-06-02	MaskOn	2019-06-02 22:11:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOff	2019-06-02 22:16:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOn	2019-06-02 22:18:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOff	2019-06-02 22:38:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOn	2019-06-02 22:39:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOff	2019-06-02 22:40:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOn	2019-06-02 22:51:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOff	2019-06-02 22:51:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOn	2019-06-02 23:41:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-06-02	MaskOff	2019-06-03 05:17:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOn	2019-05-30 19:57:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOff	2019-05-30 19:58:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOn	2019-05-30 20:00:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOff	2019-05-30 20:01:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOn	2019-05-30 20:04:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOff	2019-05-30 20:05:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOn	2019-05-30 21:45:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOff	2019-05-30 21:45:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOn	2019-05-30 21:48:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM
2019-05-30	MaskOff	2019-05-30 23:57:00	6/20/2019 3:31:02 PM	6/6/2019 4:00:26 PM

NDMC Workscope

- Need IT/IS involvement in planning
- Integrated datasets are reconciled and linked at the participant level with EDC data using unique identifiers, enabling a unified clinical data model available for central monitoring, analysis and eventually data sharing.
- Data lifecycle management, including **updates, corrections, and versioning**, can be handled through **automated workflows, audit trails, and validation rules** to ensure data integrity and regulatory compliance (e.g., 21 CFR Part 11, GCP).

Challenges and Take Home

Maintain	Maintain frequent, structured communication with centralized data providers to quickly resolve issues
Anticipate	Anticipate that technology and security changes (e.g., credentials, protocols, platforms) can disrupt workflows—plan for adaptability
Establish out	Establish out-of-band data load processes to handle corrections and missing data when standard pipelines cannot be modified
Implement	Implement a centralized knowledge repository to preserve process documentation across multi-year trials and staff changes
Retain	Retain original source data to support reprocessing, validation, and recovery during data corrections
Ensure	Ensure regular alignment between management and technical teams to maintain visibility and coordination
Have	Have a reliable contract and contact with the device provider/industry that will ensure minimal changes during the course of the trial

Challenges and Take Home

- What worked well
  - Strong collaboration and open communication with the central data management team and external device/supplier company
  - Complete data dictionary with version control process
  - Rapid identification of missing or incomplete data through close collaboration with the data management team
  - Timely data corrections and re-deliveries of new/corrected data files from device company
  - Ability to reprocess and integrate updated data efficiently
  - Continuous feedback loop improved data quality over time

Thank you

Avoiding Data Indigestion

Lisa S. Young

Sr. Director of Technology
Utah Data Coordinating Center

Research Security Officer
Sr. Director of Research IT & Cybersecurity
Office of the Vice President for Research

Society for Clinical Trials

May 20, 2026

UTAH DCC
DATA COORDINATING CENTER

DISCLOSURE STATEMENT

No traditional disclosures.

I used a lot of AI (models used: Anthropic's Opus, OpenAI's GPT 5.5, Manus Max, Google Gemini 3.5)

Empowering Innovation

Through Comprehensive Research Solutions

Platform

HYBRID on-prem + AWS for scale and availability

SECURE research enclaves with governed workloads

AI-ready landing zones for advanced research

People

Faculty LEADERS in clinical, medical, epidemiology, and statistics

Integrated TEAMS across biometrics, IT, and clin-ops

Positioning

Among the world's LARGEST academic research orgs

We PARTNER with pharma, biotech, and medtech on FDA-regulated trials, real-world data, and industry-sponsored studies to speed translation and impact.

Services

TECHNICAL Data engineering, pipelines, APIs, portals, UX

SCIENTIFIC Biostatistics, informatics, analytics, data management

CLIN-OPs Project & site management, protocol dev

Compliance

COMPLIANCE-aligned: FISMA Moderate, NIST 800-171

AUDIT-ready: GCP, 21 CFR § 11

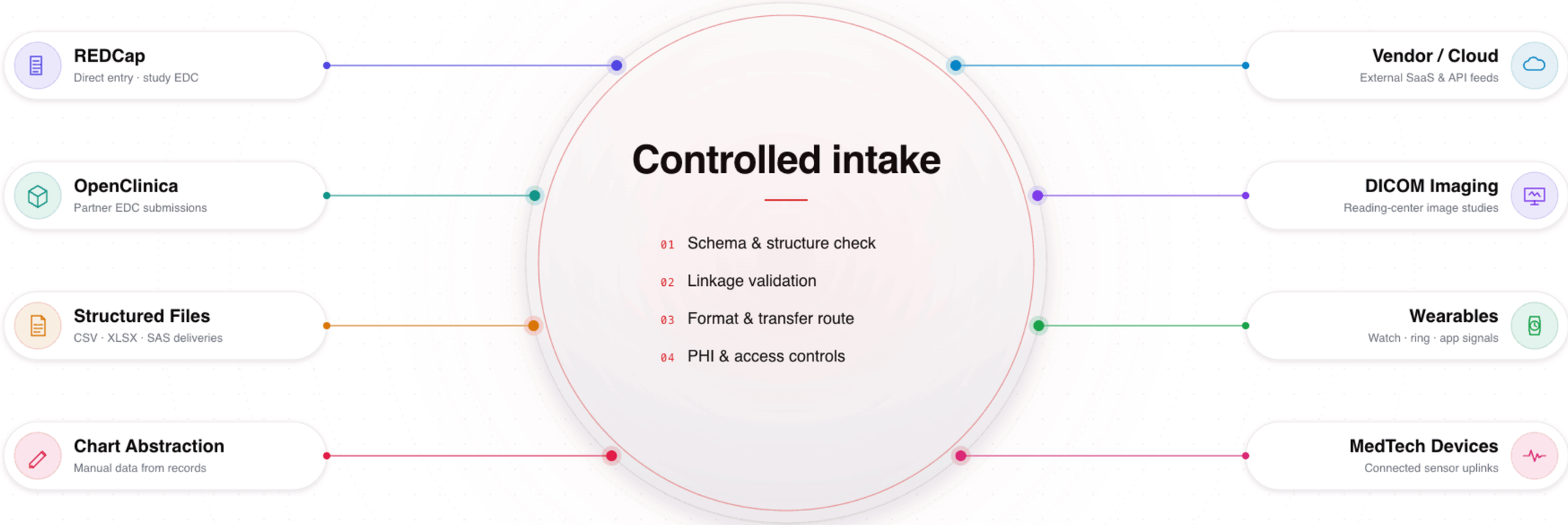
EVIDENCE-driven control framework

QUALITY-by-design: SOPs, CAPAs, trainings, reviews

HEALTH
UNIVERSITY OF UTAH

UTAH DCC
DATA COORDINATING CENTER

Many Sources, One Controlled Intake



Receipt is not readiness.

A file can arrive on time and still be unusable when structure, linkage, timing, or transfer assumptions don't line up.

Site, vendor, device

Sites & labs
Vendors & CROs
Wearables & apps
Reading centers

Transfer

SFTP
API
Portal upload
File drop

DCC intake check

Structure
Linkage
Timing
Change

Study database

EDC
Warehouse
Analytic marts

Reports & analysis

DSMB reports
Dashboards
Statistical analysis

External Data Intake Checklist

#	ITEM
1	**Site / vendor readiness** • Source owner identified • Transfer contact known • Intake expectations shared • Test file requested
2	Expected structure + data dictionary
3	File format + transfer route
4	Participant linkage
5	Timing + cadence
6	Change notification
7	Monitoring + escalation

Small upstream changes, large downstream consequences.

To a person, the same spreadsheet. To the loader, a different structure.

A EXPECTED FILE

SUBJECT ID	VISIT	DRAW DATE	HEMOGLOBIN	PLATELETS
1001	Month 6	2026-05-01	13.2	210
1002	Month 6	2026-05-01	12.8	185

B RECEIVED FILE

SUBJECT ID	VISIT	DATE DRAWN	HGB	PLATELETS
1001	Month 6	5/1/26	13.2 g/dL	210
1002	Month 6	5/1/26	12.8 g/dL	185

COLUMN RENAME
~~Draw Date~~ › **Date Drawn**

VARIABLE RENAME
~~Hemoglobin~~ › **Hgb**

DATE FORMAT
~~2026-05-01~~ › **5/1/26**

NUMERIC › TEXT
~~13.2~~ › **13.2 g/dL**

External Data Intake Checklist

#	ITEM
1	Site / vendor readiness
2	**Expected structure + data dictionary** • Column names match • Units stay consistent • Date format defined • Dictionary reflects current file
3	File format + transfer route
4	Participant linkage
5	Timing + cadence
6	**Change notification** • Site alerts DCC before changes • Dictionary updated if approved
7	Monitoring + escalation

Same data. Different file type. Different reality.

CSV vs Excel on the same tab

CSV (lab_results_2026_05_01.csv)

What people assume

	A	B	C	D	E	F
1	SampleID	PatientID	Test Date	Test Code	Result	Units
2	000123	P-001	2026-04-30	GLU	98	mg/dL
3	000124	P-001	2026-04-30	A1C	5.6	%
4	000125	P-002	2026-04-30	GLU	105	mg/dL
5	000126	P-003	2026-04-30	CHOL	180	mg/dL
6	000127	P-003	2026-04-30	HDL	55	mg/dL
7	000128	P-003	2026-04-30	LDL	101	mg/dL
8	000129	P-004	2026-04-30	GLU	89	mg/dL
9	000130	P-005	2026-04-30	A1C	6.1	%
10	000131	P-005	2026-04-30	CHOL	195	mg/dL
11	000132	P-006	2026-04-30	GLU	110	mg/dL
12						

What's different

- Multiple tabs (data on Tab 2)
- Header row shifted down
- Hidden column exists in Excel
- Dates auto-converted
- Leading zeros stripped from IDs
- Results are formulas, not fixed values

Excel (lab_results_2026_05_01.xlsx) – Sheet1

What the system actually sees

	A	B	C	D	E	F	G
1	Lab Results Export						
2	Generated: 05/01/2026 08:12						
3							
4	Sample ID	Patient ID	Test Date	Test Code	Result	Units	Comments
5	123	P-001	4/30/2026	GLU	=100-2	mg/dL	
6	124	P-001	4/30/2026	A1C	=5.6	%	
7	125	P-002	4/30/2026	GLU	=105	mg/dL	
8	126	P-003	4/30/2026	CHOL	=180	mg/dL	
9	127	P-003	4/30/2026	HDL	=55	mg/dL	
10	128	P-003	4/30/2026	LDL	=100+1	mg/dL	
11	129	P-004	4/30/2026	GLU	=90-1	mg/dL	
12	130	P-005	4/30/2026	A1C	=6.1	%	
13	131	P-005	4/30/2026	CHOL	=195	mg/dL	

README | Data | Lookup

- One tab
- Header in row 1
- IDs keep leading zeros
- Dates as text (YYYY-MM-DD)
- Values are fixed
- No hidden columns or formatting

- Multiple tabs (data on "Data" tab, not first tab)
- Header row is on row 4, not row 1
- Hidden column (Comments) exists
- Dates auto-converted to Excel date format
- Leading zeros removed from IDs
- Results are formulas, not fixed values

Changing the file type changes how the data are read.

CSV and Excel are not interchangeable from the intake side.

WHAT PEOPLE ASSUME	WHAT THE DCC PROCESS SEES
"It is the same rows and columns."	It's a different file type.
"Excel is easier to read."	Excel can include sheets, formulas, hidden columns, merged cells, formatting.
"The values are the same."	IDs, dates, and numbers may be interpreted differently.

PREVIOUSLY

`lab_results_2026_05_01.csv`

NOW ARRIVING AS

`lab_results_2026_05_01.xlsx`

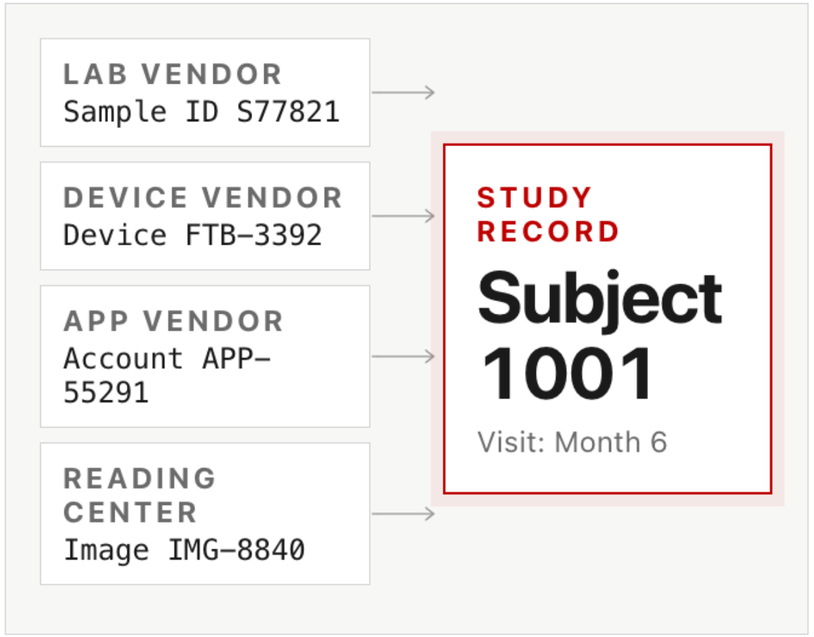
- Workbook now has multiple tabs; correct data may be on tab 2
- Header row shifted; formulas instead of fixed values
- Dates auto-converted and leading zeros stripped from IDs

External Data Intake Checklist

#	ITEM
1	Site / vendor readiness
2	Expected structure + data dictionary
3	**File format + transfer route** • File type approved • Correct sheet confirmed • Header row fixed • Transfer path stable
4	Participant linkage
5	Timing + cadence
6	**Change notification** • Format changes reviewed first • Loader tested before production
7	Monitoring + escalation

The value can be right, and still linked to the wrong person.

One participant; five different IDs. The DCC has to reconcile them into one study record.



COMMON LINKAGE FAILURES

01 Screening ID used where randomization ID is expected

02 Participant ID mistyped at the site

03 Participant changes device mid-study

04 App deleted and reinstalled, new account ID

05 Duplicate participant record created across systems

06 Visit labels differ across vendors and EDC

External Data Intake Checklist

#	ITEM
1	Site / vendor readiness
2	Expected structure + data dictionary
3	File format + transfer route
4	**Participant linkage** • Subject ID reconciled • Device ID mapped • Sample ID matched • Visit window confirmed
5	**Timing + cadence** • Timestamp source known • Time zone defined
6	Change notification
7	Monitoring + escalation

Vendor/site readiness can become the critical path.

Security and risk reviews are not paperwork at the end. They are startup dependencies.

External Data Intake Checklist	
#	ITEM
1	**Site / vendor readiness** • Risk review started early • Security contact identified • DUA / access path known • Review timeline tracked
2	Expected structure + data dictionary
3	**File format + transfer route** • Transfer method approved • Test file scheduled
4	Participant linkage
5	**Timing + cadence** • Startup timeline realistic • Delays escalated early
6	Change notification
7	Monitoring + escalation

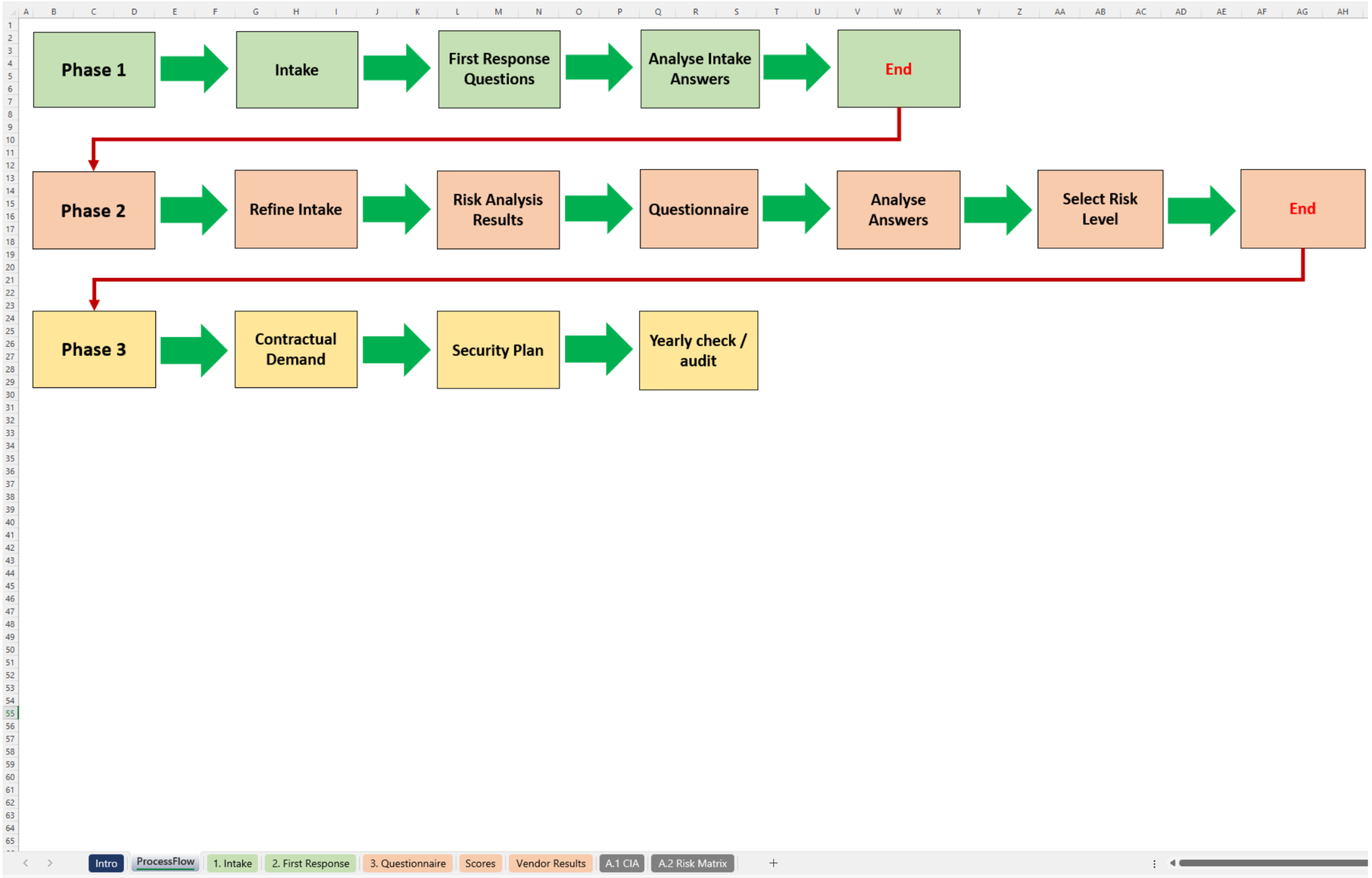
	Category	Subject	Risk level	Maturity	Questions	Answer	Justification
	Part A - Security and risk management						
1	Business Continuity	Risk	Low Risk	Level 1	Rate the compliance of your data backup facilities' locations with relevant legal and regulatory requirements	Fully Compliant	
3	Business Continuity	Contract	Med Risk	Level 1	Do you have a protocol to notify us about changes in backup providers or locations?	Yes	
4	Governance	Risk	Low Risk	Level 1	Rate the adequacy of your response and mitigative actions to any past security breaches	Slightly Adequate	
5	Supply Chain Management	SLA	Med Risk	Level 1	Can you restrict our access to data under certain circumstances, such as non-payment of fees?	Yes	
6	Supply Chain Management	SLA	Med Risk	Level 1	Will you assign a Single Point of Contact for 24/7 support?	Yes	
7	Business Continuity	SLA	Low Risk	Level 2	Is it possible for us to maintain a local backup of our data in addition to your backup solutions?		
8	Change Control	SLA	Med Risk	Level 2	Rate the effectiveness of your policies for managing risks related to changes in applications, APIs, and network/system components		
9	Governance	SLA	Med Risk	Level 2	Does your SLA involve any transfer of intellectual property or ownership rights of our data and systems?"		
10	Governance	Risk	High Risk	Level 2	Rate the impact on security and integrity of our data when you utilize cloud services from other providers		
11	Governance	Risk	Med Risk	Level 2	Do you have formal agreements with subcontractors to ensure they comply with our security requirements		
12	Governance	Risk	Med Risk	Level 2	Rate the security level of your backup storage locations		
13	Infastructure & Virtualization	SLA	Low Risk	Level 2	Do you employ file integrity monitoring and network intrusion detection systems?		
14	Encryption & Key Management	Risk	Low Risk	Level 3	If data is not stored in encrypted form, can we use our encryption solutions to safeguard it		
15	Governance	SLA	Med Risk	Level 3	Rate the efficiency of data retrieval in the event of your business cessation		
16	Governance	SLA	Med Risk	Level 3	Are you agreeable to participating in a cloud risk assessment upon our request?		
17	Governance	Risk	Critical Risk	Level 3	Can you provide your disaster recovery and business continuity plan?		
18	Governance	Technical	High Risk	Level 3	Rate the effectiveness of your controls to prevent and monitor unauthorized software installations		
19	Infastructure & Virtualization	Risk	Med Risk	Level 3	Do you regularly review audit logs for security events using automated tools?		
20	Identity & Access Management	SLA	Med Risk	Level 4	Is there an automated workflow integrating your HR systems with access control systems for timely revocation of access rights?		
21	**Part B – Compliance**						
22	Data Security Information lifecycle	SLA	Med Risk	Level 1	How would you rate your policies and technical measures for secure data disposal from storage media?		
23	Governance	Risk	Med Risk	Level 1	Do your server locations comply with data sovereignty and privacy regulations?		
24	Governance	Risk	Low Risk	Level 1	Are any employees who may access our data U.S. citizens, and does this impact data access and control?		
25	Governance	Risk	Low Risk	Level 1	How would you rate the guarantees regarding the security and integrity of data in your custody?		
26	Governance	SLA	Med Risk	Level 1	What guarantees can you provide regarding the security and integrity of our data in your custody?		
27	Governance	Privacy	Med Risk	Level 1	Do you agree to comply with our Data Processing Agreement, recognizing us as the data controller and your organization as the data processor?		
28	Governance	Contract	Med Risk	Level 1	Do you have processes to ensure subcontractors adhere to the same security requirements?		
29	Governance	Contract	Med Risk	Level 1	Do you conduct and share findings of annual penetration tests and vulnerability scans?		

Yes
No

Intro | ProcessFlow | 1. Intake | 2. First Response | 3. Questionnaire | Scores | Vendor Results | A.1 CIA | A.2 Risk Matrix

HEALTH
UNIVERSITY OF UTAH

UTAH DCC
DATA COORDINATING CENTER


Phase 1
Intake
First Response Questions
Analyse Intake Answers
End
Phase 2
Refine Intake
Risk Analysis Results
Questionnaire
Analyse Answers
Select Risk Level
End
Phase 3
Contractual Demand
Security Plan
Yearly check / audit
Intro
ProcessFlow
1. Intake
2. First Response
3. Questionnaire
Scores
Vendor Results
A.1 CIA
A.2 Risk Matrix

Minimum intake checklist for external data.

01 Who is sending the data?
Site, vendor, lab, device platform, reading center, registry, or app.

02 Is the site / vendor ready?
Security, privacy, access, DUA, IRB, and operational contacts are resolved.

03 What structure is expected?
Field names, data types, units, allowed values, and missing-value codes are documented.

04 Is the data dictionary current?
It reflects the file being sent now, not an older version.

05 What file format is approved?
CSV, XLSX, XML, SAS, API, or another format is specified and tested.

06 How will the data arrive?
SFTP, API, portal, cloud folder, or another route is documented.

07 How are participants linked?
Subject, device, sample, visit, accession, and account IDs reconcile.

08 What timing is expected?
Frequency, latency, timestamps, time zones, and late-data handling are defined.

09 What happens when a file fails?
Files are checked before load; exceptions quarantined; owners alerted.

10 What happens when something changes?
Structure, format, export logic, or vendor algorithm changes trigger review.

RECEIPT ≠ READINESS | External data need an intake check before they become study data.

External Data Intake Checklist

#	ITEM
✓	Site / vendor readiness
✓	Expected structure + data dictionary
✓	File format + transfer route
✓	Participant linkage
✓	Timing + cadence
✓	Change notification
✓	Monitoring + escalation

UTAH DCC
DATA COORDINATING CENTER

thank you.