PharmaSUG 2015 - Paper BB02

A New Era: Open access to clinical Trial Data - A case study

Aruna Kumari Panchumarthi, Novartis Pharmaceuticals Corporation, EH, NJ-USA
Jacques Lanoue, Novartis Pharmaceuticals Corporation, East Hanover, NJ, USA

ABSTRACT

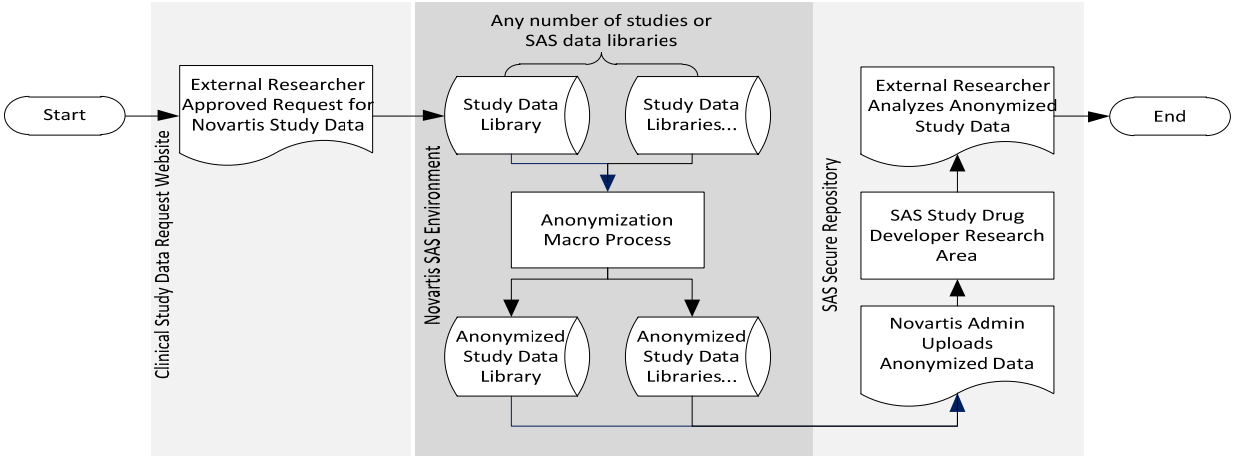
Access to the underlying (patient level) data that are collected in clinical trials provides opportunities to conduct further research that can help advance medical science or improve patient care. This helps ensure the data provided by research participants are used to maximum effect in the creation of knowledge and understanding. Researchers can use anonymised patient level data and supporting documents from clinical studies to conduct further research. This paper will present: *Overview of Data Sharing Process. *Challenges in Data Sharing. *Challenges around Anonymization. *Define new business process.

INTRODUCTION

Pharmaceutical companies, academic researchers, and government agencies such as the Food and Drug Administration and the National Institutes of Health all possess large quantities of clinical research data. If these data were shared more widely within and across sectors, the resulting research advances derived from data pooling and analysis could improve public health, enhance patient safety, and spur drug development. Data sharing can also increase public trust in clinical trials and conclusions derived from them by lending transparency to the clinical research process. Much of this information, however, is never shared. Retention of clinical research data by investigators and within organizations may represent lost opportunities in biomedical research. Despite the potential benefits that could be accrued from pooling and analysis of shared data, barriers to data sharing faced by researchers in industry include concerns about data mining, erroneous secondary analyses of data, and unwarranted litigation, as well as a desire to protect confidential commercial information. Academic partners face significant cultural barriers to sharing data and participating in longer term collaborative efforts that stem from a desire to protect intellectual autonomy and a career advancement system built on priority of publication and citation requirements. Some barriers, like the need to protect patient privacy, pre- sent challenges for both sectors. Looking ahead, there are also a number of technical challenges to be faced in analyzing potentially large and heterogeneous datasets.

OVERVIEW OF DATA SHARING PROCESS

Patient-level data collected in Novartis clinical trials will be anonymized according to the standards set forth in this document. These standards will ensure compliance with current privacy laws and regulatory guidance while allowing data to be shared with researchers. There are a number of data elements enumerated in the “Privacy Rule” under the Health Insurance Portability and Accountability Act (HIPAA) of 1996 and other guidance from European General Data Protection Regulation which can be used to identify individuals. The process of anonymizing can be thought of as permanently removing the ability to use any of these elements to identify individual participants. Direct and indirect identifiers are removed thereby making it unlikely to allow any individual to be identified by combining data. Adherence to the framework of these standards will minimize the risks of encroaching on the privacy and confidentiality of research participants. Novartis is committed to sharing clinical trial data with external researchers and has been doing so voluntarily for several years through its own web portal.



GENERAL APPROACH

WHAT DO WE SHARE?

Upon approved requests, the following data and accompanying trial documentation will be shared with qualified external researchers when available. This document will focus on the last two points below.

- ORIGINAL PROTOCOL AND ANY AMENDMENTS
- ORIGINAL DOCUMENTATION AND AMENDMENTS THAT ARTICULATE STATISTICAL METHODOLOGY
- CSR (REDACTED) APPENDICES
- ANNOTATED CRF
- DATASET SPECIFICATIONS
- ANONYMIZED RAW STUDY DATASETS – COLLECTED DATA FROM EACH PATIENT IN THE STUDY
- ANONYMIZED ANALYSIS-READY DATASETS – DATA USED FOR ANALYSIS

ANONYMIZATION PROCESS

REMOVING PERSONALLY IDENTIFIABLE INFORMATION (PII)

There are 18 identifiers to be removed from the datasets (and related documentation) as described in (HIPAA) CFR – Title 45: Public Welfare, Subtitle A §164.514. The identifiers to be removed are:

- Names
- All geographic subdivisions smaller than a state including:
  Street address, City, County, Precinct, Zip Code and Geocodes
  Except for the initial 3 digits of a zip code if:
    - The geographic area formed by combining all zip codes with the same three initial digits contains more than 20,000 people and
    - The initial three digits of a zip code for all such geographic units containing 20,000 or fewer people are changed to 000.
- All elements of dates (except year) for dates directly related to an individual, including birth date, admission date, discharge date, date of death; and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
- Telephone numbers
- Fax numbers
- Electronic mail addresses
- Social security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers and serial numbers, including license plate numbers
- Device identifiers and serial numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric identifiers, including finger and voice prints
- Full face photographic images and any comparable images
- Any other uniquely identifying number, characteristic, or code

This will be used as a framework for defining the Novartis anonymization standards, discussed in the following sections.

IDENTIFIERS

Change the real value to a de-identified value in a consistent manner so that the value in one instance of the variable is consistent with the value in the same variable across other datasets. This does not limit but includes PK datasets and central lab data. Extension studies use the same new identifiers as were used in the initial study to preserve the links between studies. This also applies to long-term follow-up studies where separate reports are published.

- The investigator number is re-coded or set to blank for each investigator. The investigator name is set to blank or dropped from the dataset.
- Each participant is given a new subject identifier.
- Each center is given a new identifier. Trials containing one or more center with <10 patients will need to be dealt with on a trial by trial basis. Aggregation of centers can be considered or possibly dropping center.

FREE TEXT VERBATIM TERMS

Information in a descriptive free text verbatim term may compromise a participant's anonymity.

- Free text verbatim terms are not included and are set to blank or dropped from the following datasets:
  - Adverse Events
  - Medications
  - Medical History
  - Other specific verbatim free text
- All standard dictionary coded terms will be retained.

DATE OF BIRTH

Information relating to a research participant's date of birth and identification of specific ages above 89 may compromise anonymity.

Date of birth is dropped and ages above 89 are aggregated into a single category of "90 or older". (See section 5 example)

OTHER DATES

Specific dates directly related to a research participant may compromise a research participant's anonymity.

- A random offset per study, is generated and applied to all dates. All original dates are replaced with the new dummy dates so that the relative times between dates are retained.
- This date offset will be generated at the study level pushing dates into the future.
- Do not retain any seasonal information.

Example: If the original reference date was 01APR2008 and the date of death was 01MAY2008, a random offset is generated (in this case 74,916 days). Dummy dates are then calculated using this offset of 74,916 days.

	Original Date	New Date	
Reference Date	01APR2008	13May2213	Apply offset = 74,916 days
Date of Death	01MAY2008	12JUN2213	Apply offset = 74,916 days
Relative Time of Death	30 days	30 days	

OTHER PII

Other data elements that contain PII are removed. For example:

- Information from variable names e.g. lab names may contain location information
- Investigator comments may be used to identify a subject
- Genetic data will be not be shared at all
- Exploratory Biomarker data outside the primary and key secondary endpoints and laboratory data
- Also excluded will be case narratives, documentation for adjudication and imaging data (e.g. x-rays, MRI scans)

REMNANTS

After anonymization, there is no information available that will allow us to recreate the original datasets from the anonymized data. This includes but is not limited to the following:

- Any transactional copies of anonymized datasets
- De-identification tables (links from original variable to new anonymized variable)
- QC output datasets
- Any Log or LST files
- The seed utilized for random number generation

The anonymized datasets are stored separately from the original datasets in the Novartis systems.

Result of Anonymization process

- **Example**

Study data example on top and anonymized data on bottom after ***modes of anonymization*** were applied.

Study Data

Center ID	Investigator ID	Investigator name	Subject number	Date of birth	Age (yrs)	AE start date	AE end date	Verbatim term	Preferred term
T1230	279T344	Dr Smith	2002	08Aug1954	57	29DEC2010	27JAN2011	HEADACHE	Headache
T1230	279T344	Dr Smith	2002	08Aug1954	57	10JAN2011	06APR2011	BRONCHITIS	Bronchitis
T1230	279T344	Dr Smith	2004	09Aug1919	92	25MAR2011	12AUG2011	COLD	Nasopharyngitis
T1230	279T344	Dr Smith	2004	09Aug1919	92	28MAR2011	31MAR2011	FLU	Influenza
T1230	279T344	Dr Smith	2004	09Aug1919	92	01MAR2011	15MAY2011	PAIN	Pain
G5670	348G224	Dr Jones	2010	09Aug1947	64	14OCT2010	20OCT2011	ACHE NOS	Pain
G5670	348G224	Dr Jones	2010	09Aug1947	64	24MAY2011		BRONCHIAL INFECTION	Bronchitis
G5670	348G224	Dr Jones	2010	09Aug1947	64	01MAR2011	15MAR2011	CHRONIC PAIN	Pain
TRANSLATE	TRANSLATE	MISSING	TRANSLATE	DROP	AGEINT	DATE	DATE	MISSING	NONE

Anonymized Data

Center ID	Investigator ID	Investigator name	Subject number		Age (yrs)	AE start date	AE end date	Verbatim term	Preferred term
Xnn10	nnnXn10		Ay12		57	16FEB2093	17MAR2093		Headache
Xnn10	nnnXn10		Ay12		57	28FEB2093	25MAY2093		Bronchitis
Xnn10	nnnXn10		Eb65		90 or Older	13MAY2093	30SEP2093		Nasopharyngitis
Xnn10	nnnXn10		Eb65		90 or Older	16MAY2093	19MAY2093		Influenza
Xnn10	nnnXn10		Eb65		90 or Older	19APR2093	03JUL2093		Pain
Xnn11	nnnXn11		Nz97		64	02DEC2092	08DEC2093		Pain
Xnn11	nnnXn11		Nz97		64	12JUL2093			Bronchitis
Xnn11	nnnXn11		Nz97		64	19APR2093	03MAY2093		Pain

MODES OF ANONYMIZATION:

- DROP

VIEWTABLE: Rchwork.Input

	TEST_VAR1	TEST_VAR2
1	DATA	data
2	DATA	data
3	DATA	data

VIEWTABLE: Rchwork.Output

	TEST_VAR2
1	data
2	data
3	data

- MISSING

VIEWTABLE: Rchwork.Input

	TEST_VAR1	TEST_VAR2
1	DATA	data
2	DATA	data
3	DATA	data

VIEWTABLE: Rchwork.Output

	TEST_VAR1	TEST_VAR2
1		data
2		data
3		data

- TRANSLATE

VIEWTABLE: Rchwork.Input

	TEST_VAR1
1	SUBJ01
2	SUBJ01
3	SUBJ04

VIEWTABLE: Rchwork.Output

	TEST_VAR1
1	XXXXnn
2	XXXXnn
3	XXXXnn

- DATE

VIEWTABLE: Rchwork.Input

	TEST_DT
1	06-25-1985
2	08-16-1986
3	12-31-1991

VIEWTABLE: Rchwork.Output

	TEST_DT
1	07-26-2085
2	09-17-2086
3	01-01-2091

- AGEINT

VIEWTABLE: Rchwork.Input

	TEST_AGE
1	60
2	96
3	55

VIEWTABLE: Rchwork.Output

	TEST_AGE
1	60
2	90 or older
3	55

- NONE (straight copy of the variable)

** By default in the macro, if no mode is defined for a variable, the variable is dropped*

CHALLENGES IN DATA SHARING

Many trials do not use the standard Informed Consent Form (ICF) but use the site ICF or their own version of an ICF - Need to ensure anonymization is inserted into these ICF's without exception. Informed Consents have changed over time and may be restricting the use of the data only for the study in question – additionally individual Ethics Committees can propose alterations to ICF. It is easy to anonymize data but to create an anonymized database that maintains the within study and within patient data relations needed for analyses is more difficult.

HIPAA (only applies to US) provided some high level considerations but additional definitions and specifications are needed in EU. Eg: height/weight; dealing with small subgroups (center, gender); handling dates etc.

CHALLENGES IN DATA ANONYMIZATION

The raw and analysis ready datasets will be anonymized where all personally identifiable information (PII) will be removed or replaced. Subject identifiers will be recoded. Free text will be removed. Date of birth will be dropped; age will be categorized. Dates will be offset to a point into the future. There will be no way to undo and recreate the original data once it is anonymized.

NEW BUSINESS PROCESS

Novartis defined a new business process is to describe the end-to-end procedure of taking final clinical study SAS data sets, running the anonymization macros on them to create anonymized clinical study data sets and transferring them to a SAS Secure Repository that enables external researchers to access and use these anonymized data subject to their approved research proposal and data sharing agreement.

REFERENCE LIST

Guidance on De-identification of Protected Health Information – US
http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule
http://privacyruleandresearch.nih.gov/pdf/HIPAA_Booklet_4-14-2003.pdf

European Union General Data Protection Regulation
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:008:0001:0022:en:PDF

ACKNOWLEDGMENTS

We'd like to acknowledge and thank our Global Programming Head Guillaume Breton for reviewing this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Name: Aruna K Panchumarthi
Organization: Novartis Pharmaceuticals
Address: 1 Health Plaza
City, State ZIP: East Hanover, NJ.
Email: Kumari.sai@novartis.com

Name: Jacques Lanoue
Organization: Novartis Pharmaceuticals
Address: 1 Health Plaza
City, State ZIP: East Hanover, NJ.
Email: Jacques.Lanoue@novartis.com