An Introduction to:
Good Data Management Practice (GCDMP),
Good Documentation Practice &
Good Storage / Archiving Practice
Acknowledgment

Acknowledgment to Ms Teo Jau Shya and Chun Geok Ying for preparation of the core contents of this presentation.
Objective

- Know what are data
- Understand the importance of data management
- Be able to understand key data and process of collecting the data
- Know best practice of reporting data in CRF
- Awareness of Good documentation practice
- Understand how to store and archive data and the related research documents
Good Clinical Data Management Practice - why is it important?

• Data are the most important product of clinical research = oxygen in science
• The ability to record, store, manipulate, analyze, and retrieve data is critical to the research process
• To provide consistent, accurate & valid data
• To support the accuracy of the final conclusion & report
Consequences of bad data management

– A small amount of data errors, even if the effect is minimal on scientific conclusions, the effect of discovered errors in the data on public perception & external acceptance of the results can be profound.

Example:

– In 2006, Japanese researchers reported\(^1\) a technique for creating cells that have the embryonic ability to turn into almost any cell type in the mammalian body — the now-famous induced pluripotent stem (iPS) cells.
Stimulus-triggered reprogramming into pluripotency

Haruko Obokata, Teruhiko Wakayama, Niwa, Masayuki Yamato & Charles A. Vacanti

Affiliations | Contributions | Corresponding authors

Bidirectional developmental potential in reprogrammed cells with acquired pluripotency

Haruko Obokata, Yoshiki Sasai, Hitoshi Niwa, Mitsutaka Kadota, Munazah Andrabi, Nozomu Takata, Mikiko Tokoro, Yukari Terashita, Shigenobu Yonemura, Charles A. Vacanti & Teruhiko Wakayama

TOKYO—Hailed two months ago as a breakthrough in Japanese stem-cell research, a string of retractions has followed.

Affiliations | Contributions | Corresponding authors

Nature 505, 676–680 (30 January 2014) | doi:10.1038/nature12969
Received 10 March 2013 | Accepted 20 December 2013 | Published online 29 January 2014
Retraction (July, 2014)
Stem cell debacle déjà vu
by Michael Cook | 6 Apr 2014 |
tags: research ethics, research misconduct, stem cell ethics, stem cells

Once again, a major advance in stem cell science has been talarations of fraud. A leading research centre in Japan, the has apologised for “research misconduct” by a young scientist, paper in Nature in January which described an exciting new producing pluripotent stem cells.

Public misled by distorted results into drawing conclusion

But, because of poor data management and the failure to properly label samples in the laboratory, “it’s impossible to know exactly where it came from,” he says.

The ease of the method astonished stem cell biologists, but no one was able to replicate the results and careful scrutiny of the paper soon revealed some troubling problems. It revived painful memories of the notorious Hwang Woo-suk and the fraudulent stem cell papers he published in 2005 in the journal Science.

Haruko Obokata now stands accused of misusing images to support the creation of what she and her colleagues termed “stimulus-triggered activation of pluripotency” (STAP) cells. She admits that there were mistakes in her paper, but denies the charges of misconduct. She also insists that the method does work.

A six-member committee set up by the RIKEN Institute was asked to investigate six errors in the paper. Four were found to be innocent mistakes, but two were branded “misconduct”. “Dr Obokata’s actions and sloppy data management lead us to the conclusion that she sorely lacks, not only a sense of research ethics, but also integrity and humility as a scientific researcher,” was their brutal conclusion. Nature is conducting its own investigation.

The president of RIKEN, Nobel laureate Ryoji Noyori, issued an apology both for Obokata’s alleged misconduct and for the carelessness of her co-authors. If the misconduct is confirmed after the appeals process, he said that the paper should be retracted and that the authors would be disciplined.
DATA SEQUENCE

Protocol Design → Case report form design → Database design

Data cleaning, verification and dataset locked → Data entry → Data collection

Statistical analysis → Report writing → publication
Data Collection Tool - Case report form (CRF)

- Data from a clinical research is collected in a CRF.
Development of a Data management strategy

• Purpose: to turn Data into knowledge ➞ translate relevant activities within protocol into data

• Planning stage-How to design data collection tool/CRF:
  – Identify data sources ➞ Primary or Secondary sources?
  – Clearly define all data to be collected before seeing first subject
    • Data field clearly define and consistent throughout including base units
    • Ie: Blood pressure: Systolic BP (mmHg) + Diastolic BP (mmHg); Height (cm)
    • Store raw data fields: ie: Ht and Wt instead of BMI
  – Identify a list of possible confounders: to adjust for final data analysis
  – Identify data collection strategy: data structured? Data is available from patient’s records? Need to interview subjects?
Development of a Data management strategy

• Planning stage (continue)
  – Using questionnaire to collect data: Validated-in English and local language? Any licensing issue?
  – If outcome measures (ie: Size of tumour) is derived from X-ray, ECG: identify means to store those high density data
  – Finalized data collection tool → state version number and effective date
  – **Quantitative** studies: reduce free-form text data (ie: narrative): difficult to analyse
  – **Quantitative** studies: unstructured data, transcribe into text: how to capture data? Recording? Video taping?
  – Identify who fills up the CRF and conduct training
  – Have a standard operating procedure /manual
Sample of CRF

<table>
<thead>
<tr>
<th>Demographics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject UWHC Medical Record Number</strong>: [ ] [ ] [ ]</td>
</tr>
<tr>
<td><strong>First Name</strong>: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]</td>
</tr>
<tr>
<td><strong>Middle Name (or initial)</strong>: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]</td>
</tr>
<tr>
<td><strong>Last Name</strong>: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ]</td>
</tr>
</tbody>
</table>
| **Birthday**: [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ }
DATA SEQUENCE

Protocol Design → Case report form design → Database design

Data cleaning, verification and dataset locked

Data entry → Data collection

Statistical analysis → Report writing → publication
Development of a Data management strategy

• Implementation stage:
  – Build database/spreadsheet to receive data extracted from CRF
  – For categorical variables: code data in a consistent manner (ie: use similar codes)
    • Ie: use of medical coding: ie: MeDRA for coding of adverse events
    • For missing data: avoid using “O”
    • Purpose: to facilitate statistical analysis
  – Include data validation plan: to check any discrepancies, transcription errors or calculation for derived variables like BMI, age
<table>
<thead>
<tr>
<th>Subject ID</th>
<th>Birthdate (dd/ mm/ yy)</th>
<th>Gender (0=male, 1=female)</th>
<th>Ethnicity (0= Hispanic, 1= non-hispanic)</th>
<th>Race (0=American indian/alaska native, 1= Asian, 2= Black African American, 3=native hawaiian, 4= white or caucasian)</th>
<th>Height (cm)</th>
<th>Weight (kg)</th>
</tr>
</thead>
</table>
GOOD CLINICAL DATA MANAGEMENT PRACTICE

Why need a database/spreadsheet and why not direct data entry into SPSS or any statistical software?

GOOD DATA MANAGEMENT PRACTICE
Spreadsheets vs Database

• Databases are used to **protect, store, and retrieve data**
  – Eg: Microsoft access, OpenClinica (an open source database), etc

• Databases are **safer**. Excel, for example, does everything in memory, so that any **unsaved data may be lost** if your system crashes. Databases write data to the hard drive immediately.

• Databases can **handle more data**. Sure, Excel can technically handle more than 65,000 rows of data, but doing so will likely bog down even the fastest PC.

• Conclusion:
  – A spreadsheet has serious drawbacks when used for data storage, is **cumbersome to retrieve** offers little or no data validation and little or **no protection against data corruption** from well-meaning but poorly trained users.
  – Spreadsheet is better at a lot of things—displaying charts, displaying different types of data on the same worksheet.
DATA SEQUENCE

Protocol Design → Case report form design → Database design

Data cleaning, verification and dataset locked → Data entry

Data collection → Data entry

Statistical analysis → Report writing

Report writing → publication
**Characteristics of the Data**

**Quality vs Integrity**

**Quality**
- Essential Characteristic of each piece of data

<table>
<thead>
<tr>
<th>A</th>
<th>L</th>
<th>C</th>
<th>O</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
<td>E</td>
<td>A</td>
<td>Others?</td>
</tr>
</tbody>
</table>

- **Attributable** (Who wrote this?)
- **Legible** (Can I read this?)
- **Contemporaneous** (Was this recorded at the time of the results or later?)
- **Original** (Is this unaltered or copied?)
- **Accurate** (Is this a correct reflection of the conduct of the research?)

- Complete
- Consistent
- Enduring
- Available and accessible
Quality vs Integrity

Integrity

– the trustworthiness of **information**
– soundness of the body of data as a whole. In particular, the body of data should be
  • credible,
  • internally consistent, and
  • verifiable
Good Documentation Practice

• Definition
  – Concise, Legible, Accurate and Traceable
  • Concise: The document must tell the entire story & standardized
  • Legible - must be readily retrievable & readable
  • Accurate – error free, data shall be recorded as soon as possible and shall not be falsified
  • Traceable - Evidence proving that the tasks have been completed as they should be. Information on when, where, who, why and how to complete tasks
What happens when there is flawed data?

- **Lack of space between drug name and dose**
  Problem: A handwritten order for "cisplatinol (sic) 75 mg/m2" was subsequently typed as "cisplatinol75 mg/m2." The last letter (l) was misread as part of the dose. The patient received 175 mg/m2 and suffered hearing loss and acute renal failure.

- **Misplaced decimal point**
  Problem: A patient received 5 mL of fentanyl (0.25 mg or 250 mcg) instead of 0.5 mL (25 mcg) after a nurse mentally misplaced the decimal point when converting the milligram dose expressed on the label with the ordered dose in micrograms.

- **Unclear communication of orders using a felt tip pen**
  Problem: A COUMADIN (warfarin) dose duration of 2 days was misinterpreted as 7 days when the prescriber used a felt tip pen and the bottom of the numeral 2 failed to carry through to the carbon copy. Recommendation: Remind prescribers to use a **ball point pen** to write orders on multiple copy forms.

- Source
  http://www.ismp.org/Newsletters/acutecare/articles/A3Q99Action
Good Documentation Practice

• Change management / Version control
• Documentation should permit the complete reconstruction of a study
  – Record data directly, promptly and legibly in indelible ink (never pencil)
  – Initial and date all observations and any resulting changes, but do not obscure original data
  – Initial and date only work you’ve performed
  – Do not document selectively or in advance of performing the activity
Good Documentation Practice

- Do not use white-out correction fluid or tape
- Do not use ditto marks as raw data
- **Copy all heat sensitive paper** and stamp “exact copy”
Good Documentation Practice

• Documentation must allow another person to be able to accurately reconstruct what you have done

• **Keep all original observations** including those observations recorded directly into a computer

• **Sign and date all computer printouts**

• **Never back-date** anything

• **Document all deviations** with accompanying explanations

• Indicate in the **record** all applicable units and equipment used
Quiz

• Which are the following is good documentation practices
  a) date and time specific entries when medications are administered
  b) the name of the individual administering
  c) clear notation why doses are missing
  d) all of the above

• Accuracy –
  • which of the following are accurate (a) or (b):

<table>
<thead>
<tr>
<th></th>
<th>(A)</th>
<th>OR</th>
<th>(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ate 50% of the food served</td>
<td>Ate with poor appetite</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Refused medications</td>
<td>Uncooperative</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Seen crying</td>
<td>Depressed</td>
<td></td>
</tr>
</tbody>
</table>
Data entry

• To reduce transcription error: Single entry vs dual entry

• Dual entry: involves 2 person entering data into database at 2 separate time
  – Pros: most effective manner to reduce error
  – Cons: expensive and time and labour intensive

• Single entry + visual verification:
  – One person record data and review own records
  – Another person randomly select list of record and cross check against original
DATA SEQUENCE

Protocol Design → Case report form design → Database design

Data cleaning, verification and dataset locked

Data entry → Data collection

Statistical analysis → Report writing → publication
Data Validation

— Data cleaning

- Inconsistency medications, 8.4
- Missing Age, 8.2
- Invalid Date of Diagnosis, 3.2
- Missing lab results, 9
- Duplicates, 1.2

Clean data, 86

— Dataset locked after cleaning and no further changes to the data:
  * regularly backup database
EASY QUIZ

• Which of the following is a consequence of improperly collected data?

A. Provides a reliable source of data on which to base public policy.
B. Ability to answer research questions accurately.
C. Misleading other researchers to pursue fruitless avenues of investigation
D. Accurate findings result in efficient use of resources.
Core CDM Processes

• **DATA ACQUISITION**
  – Data Collection Tool Design (paper)
  – Data Collection Tool Design (electronic)

• **DATA PROCESSING**
  – Forms Processing
  – Data Entry
  – Coding

• **DATA STORAGE**
  – Database Structure Specification
  – Forms Management
  – Data Archival (paper & electronic)
Core CDM Processes

• LAB, SAFETY REPORTING & OTHER EXTERNAL DATA
  – Data Transfers & Loads
  – Database Reconciliation

• DATA QUALITY
  – Quality Control Procedure
  – Statistical Sampling
  – Quantification of Database Quality
Good Archiving Practice / Good Storage Practice
Data Storage & Archiving

• WHY Data storage is crucial to a research project?:
  – Properly storing data is a way to **safeguard** your research investment.
  – Data may need to be **accessed in the future** to **explain** or augment subsequent research or **re-create** the findings.
  – **Other researchers** might wish to **evaluate** or use the results of your research (data exchange).
  – Storing data can protect research subjects and researchers in the event of **legal allegations**.
Good Storage Practice - Storage areas

• Prevent unauthorized persons from entering
  • Consider locking cabinets to increase security and minimize dust and clutter
  • Databases should be stored in secured computers / storage media. These computers / storage media must be password-protected, and stored under lock & key (e.g. in a locked cupboard / office)

• Sufficient capacity to allow the orderly storage
• Good storage conditions - clean and dry (e.g. temperature, relative humidity), pest-control agents used
• Documentation: written instructions and records
What documents to Archive

• Definition of essential documents –
  – those which individually and collectively permit the evaluation of the conduct of a trial / research and the quality of data produced
  – Includes
    • Documents before a research (eg: approved protocol, consent forms, letter of ethics approval)
    • Documents during a research (eg: case report form)
    • Documents after a research (eg: study close out letter)
    • this includes "relevant" correspondence / communications, letters, protocol violations, research conduct, AE reporting
Data Storage & Archiving

• Archived documents should be **packed securely** in archival boxes. Staples, plastic wallets and paper clips must be removed as these will degrade the records over time.

• Paper documents must be **suitable** for long term archiving. For example, faxes or ECG results should be photocopied since the inks used may be prone to fading.

• The file should be **clearly labelled** with the name and reference number of the study, sponsor (if applicable), investigator and date to be archived until
Data Storage & Archiving
(MOH Hospital Archival Policy)

- Patient medical records are source documents
- Medical records stored in record office: can be tagged to ease retrieval
- Reference documents:
  - Jadual Pelupusan Rekod Perubatan [MOH/P/PAK/121.06 (GU) 2007]: currently in the process of updating
  - Garis panduan Pengendalian dan Pengurusan Rekod Perubatan Pesakit bagi Hospital-hospital dan Institusi Perubatan [MOH/P/PAK/199.10 (GU)]
Maintain data security

• Especially important if data involve personal identifiers
• Establish password policy
  – Control access for the PC
  – Password protect dataset within excel
  – change password every 30 days
• Assign different level of access to different personnel
  – Ie: Level 1 access: only enter record but cannot view or correct records
  – Level 2 access: enters an corrects records and views reports
Thank you