Outbreak, Surveillance, Investigation & Response (OSIR) Journal
Field Epidemiology Training Program, Division of Epidemiology
Department of Disease Control, Ministry of Public Health, Thailand
Tel: +6625903894, Fax: +6625903845, Email: osireditor@osirjournal.net, https://he02.tci-thaijo.org/index.php/OSIR

The Grammar of Science: It's All about Trustworthiness of Your Data

Jaranit Kaewkungwal*

Mahidol University, Thailand

*Corresponding author email: jaranitk@biophics.org

Received: 4 Sep 2024; Revised: 8 Sep 2024; Accepted: 9 Sep 2024
https://doi.org/10.59096/osir.v17i3.270958

As mentioned in previous article, several international standards for designing, recording and reporting research studies focus on ethical issues related to rights, safety and wellbeing of study participants and the trustworthiness of the study data.[1–4] The good clinical practice (GCP) guideline indicates that the research team should assure data generated are of sufficient quality to ensure reliable study results.[1] The guideline describes intensively on data governance issues such that there should be appropriate management of data integrity, traceability and security in order to have accurate reporting, verification and interpretation of the study information.[1] Similarly, the UK Office for Statistics Regulation notes that the framework for the code of practice for statistics is generally based on three pillars—trustworthiness, quality and value of the data.[5] There are so many jargons here—let's try to understand them.

Data Management vs. Data Governance

Data management includes tools, procedures, and methods to manage the lifecycle of data, from data initiation to data archive.[6-8] Data validation is one of the critical aspects of data management.[8] Main purpose of data validation is to ensure that the collected data conforms to predefined rules or standards by verifying its accuracy, consistency, and other metrics.[9]

Data governance is a set of processes, roles, policies, standards, and metrics that ensure the effective and efficient control, and utilization of the collected data.[7] Data governance focuses on the processes to increase value of data, while storing, manipulating and using the data, without compromising its security, integrity, or privacy.[6,8,10] Key components of data governance include policies, procedures and standards that govern data management practices regarding data ownership, data stewardship (access, maintenance, use, sharing), data protection, regulatory compliance, and data standardization (using common terminology across different systems).[7,8,11] We can say that data governance can be seen as the blueprint for constructing a new building, whereas data management is the act of construction.[6,11]

Data Quality, Data Security, and Data Integrity

Three fundamental concepts in data management are: data quality, data security and, data integrity. These terms embrace different aspects of data management.

Data quality refers to the condition of a set of values of the data, ensuring that it is fit for its intended use.[7] The purpose of data quality management is to free collected data from anomalies, inconsistencies, inaccuracies, incompleteness, repetitiveness, etc.[9,12] Data quality will streamline data analysis and produce reliable study results. We can say that data quality management consists of practices, methodologies, and tools that systematically identify, rectify, and take preventive measures against potential problems before they can disrupt the data analysis.[9,13]

Data security refers to data protection from unauthorized access and use of the collected data. The security measures should also safeguard the data from breaching or other misconducts.[12] Data security entails technologies, policies, and practices to ensure authentication of the data storage system such that only authorized persons can access to the data.

Data integrity is a broader term encompassing both data quality and security.[12] The "integrity" for data means "wholeness" and "unity".[14] According to international guidelines for research conduct, the generic definition of data integrity means the process of maintenance and assurance of the data quality over its entire data life-cycle.[1–4,14] Specifically, maintaining data integrity involves safeguarding the data against

loss, leaks and falsification, while assurance of data quality is to secure accuracy, completeness and consistency of the data at any point in its lifecycle.[12]

Data Cleaning vs. Data Cleansing

Important parts of the data management procedures after data collection include data assessment, data cleaning, and data cleansing. The purpose of data assessment is to determine if the collected data fulfills the quality standards.[15] If not, we have to perform data cleaning and/or data cleansing. These two terms are often used interchangeably, but they are actually revolved on different concepts and practices.[16] While both processes attempt to improve data quality, the choice between the two often depends on the specific requirements which may vary subject to the complexity and sensitivity of the data analysis tasks.[14]

Data cleaning primarily focuses on ensuring that the dataset is as error-free as possible before it is used for analysis. Data cleaning involves removing or correcting data that is incorrect, incomplete, duplicated, or improperly formatted. The term "data cleaning" is commonly used in academic or scientific communities with the focuses on the accuracy and reliability of data; another related term used in business settings is "data scrubbing" with the focuses on cleaning data for operational efficiency and regulatory compliance.[16] Data cleaning can be automated in the computerized system used for data quality management.[17]

Data cleansing extends beyond cleaning by adding comprehensive process of preparing data. The cleansing involve: checking data irrelevant to the study objectives, removing duplicated data, handling missing values, normalizing or standardizing data formats and structures, and ensuring the data adhere to the relevant data governance standards.[8,9,14,18] The process of data cleansing might involve cross-referencing information with external sources or employing analytic technologies to detect unanticipated patterns of incorrect data.[17,18]

Data Quality Metrics

The idea behind data integrity is to guarantee the reliability, traceability and security of data throughout all processes and systems.[19] The prominent metrics that are universally used to assess data quality in good data management practices and to evaluate document management in the good documentation practices are ALCOA and ALCOA+.[7–9,14,15,18–22]

ALCOA

ALCOA is an abbreviation of Attributable, Legible, Contemporaneous, Original, and Accurate.

A—Attributable (identifiable), being able to trace the persons involving in the processes related to data management including: generating, making corrections, deletions, additions, etc. The "attributable" can be achieved through using validated computerized system with audit trail system functions that can keep records of all activities from data entry to data archiving. For example, the validated computer system contains a journal file with records of who accessing and manipulating the data with date and time-stamped on the data records within the system. It is also captured the original values as well as modified/deleted values (Figure 1).

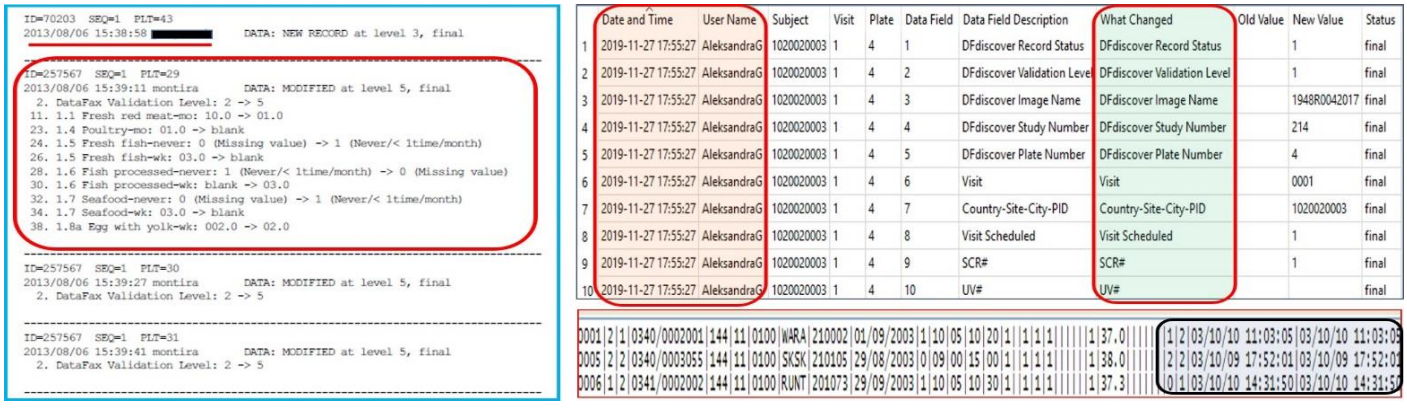

Figure 1. Examples of attributable of data

L—Legible (readable) and understandable of all information to be completed in the study. The data should also be permanent and accessible throughout the data lifecycle. For example, handwritten can be difficult to read and understand. Even though the data entry persons can guess; however, according to GCP, they cannot enter the guessing data but have to query back to the data originator/collector (Figure 2).

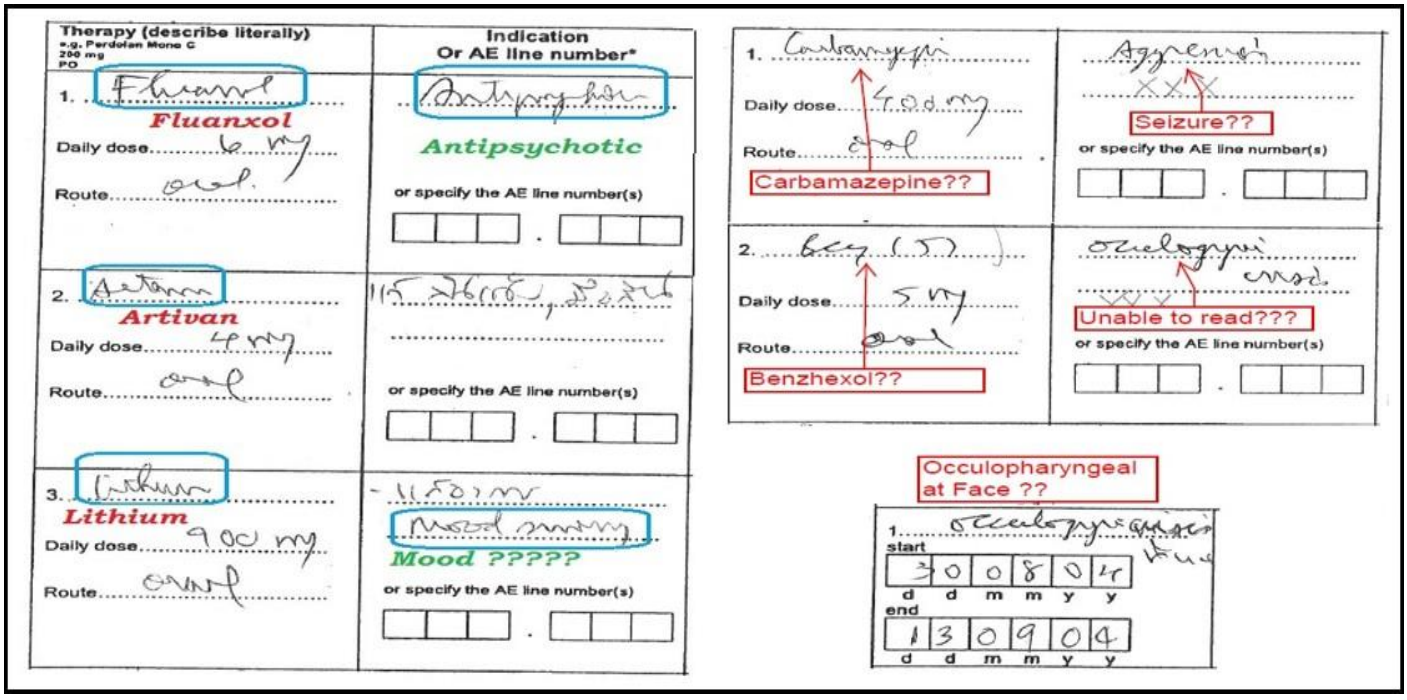

Figure 2. Examples of legible of data

C—Contemporaneous (synchronous), showing the evidence that data are simultaneously or timely documented when the actions or events are actually performed. Contemporaneous is also related to another term, timeliness, ensuring that data are up-to-date and readiness within a certain time frame. For example, as noted on the data records, the study participant enrollment dates (5 Aug 2015) are contradicted with the data submission date (4 Aug 2015). Another scenario shows the issue of unreasonable gap time between data collection date (December 2009) and data submission date (February 2010) (Figure 3).

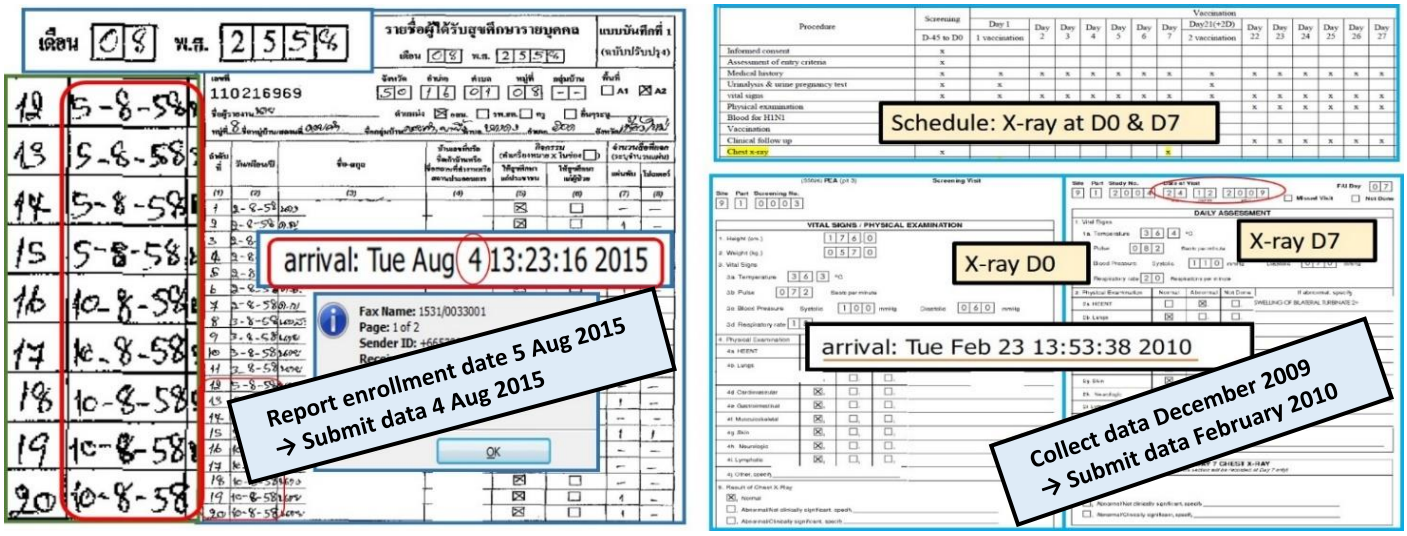

Figure 3. Examples of contemporaneous and timeliness of data

O—Original record (or certified true copy), reflecting the source of the collected information remain available in its original state. Should there be alteration made to the data/records, they should be signed and dated by an authorized person while keeping the reading of the original information. For example, according to GCP, the modification of the data should be traceable with audit trails, the data records were edited by whom and when (Figure 4).

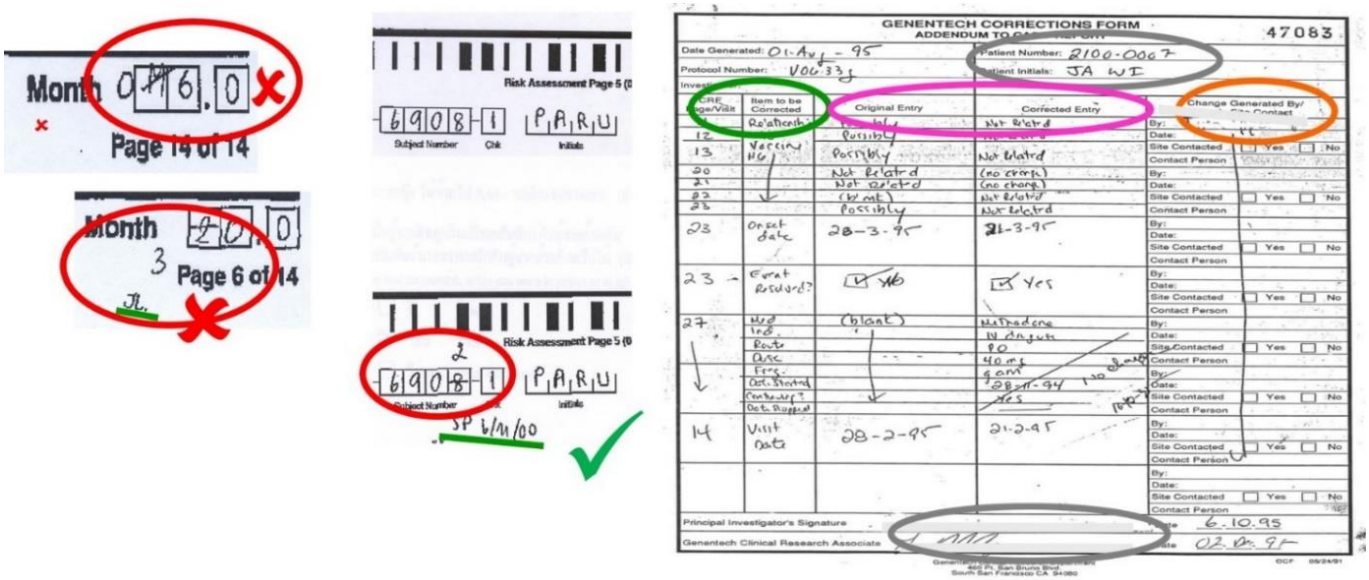

Figure 4. Example of traceable of original data

A—Accurate, verifying that the data values represent reality or are based on the agreed-upon source of truth, ensuring that the data is correct, reliable, and error free. The "accuracy" can be achieved with good automated edit check programs. In certain instances, the automated edit check between contradicting values of variables may not be possible (pre-planned), manual review by data management team is necessary particularly for key outcome variables. For example, the data on a case record form are cross-checked whether they are the same with those on the source document. Manual cross-check is performed between the medication given and the reason for such therapy (Figure 5).

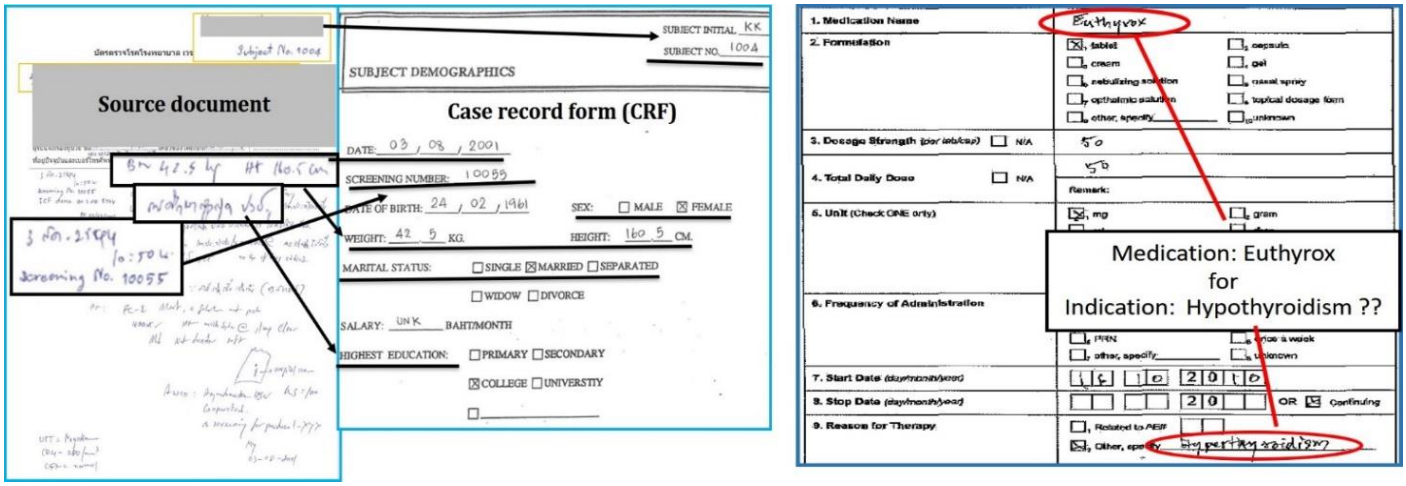

Figure 5. Examples of cross-checking of accuracy of data

ALCOA+

ALCOA+ adds four more quantifiable measurements: Complete, Consistent, Enduring, and Available.

C—Complete (whole), including all necessary data without omissions. The amount of usable or complete data should represent the sample data needed to answer the research questions as planned. Particularly, complete or meaningful data for critical variables related to primary objectives of the study should be acquired. Metadata (information about the collected data) is also important for reproducing information, if needed. For example, rather than leaving blank space for missing data, it is a good practice to assign a missing value with a specific value for each variable. Another approach is assigning "N" (none) or "ND" (not done) or "NA" (not available) for the certain variable, as applicable (Figure 5).

USUBJID	INITIALS	VISDAT	ICDAT	SEX	HEIGHT	WEIGHT	BMI	SYSBP	DIABP
03025	RK	09/APR/2014	09/APR/2014	2	.N	058	.N	130	080
03027	PN	07/MAY/2014	07/MAY/2014	1	.N	051	.N	130	073
03029	SK	07/MAY/2014	07/MAY/2014	1	.N	068	.N	128	070
03040	VS	18/AUG/2014	18/AUG/2014	2	.N	047	.N	142	063
03047	SO	20/AUG/2014	20/AUG/2014	2	.N	083	.N	123	101
03050	KS	27/AUG/2014	27/AUG/2014	2	.N	076	.N	144	077
03053	RV	03/SEP/2014	03/SEP/2014	1	.N	074	.N	132	081
03054	BR	03/SEP/2014	03/SEP/2014	2	.N	048	.N	109	069
01012	SJ	07/FEB/2014	07/FEB/2014	2	.U	067	.N	131	069
03044	VP	20/AUG/2014	20/AUG/2014	1	.U	071	.N	115	066
03049	SR	27/AUG/2014	27/AUG/2014	2	.U	051	.N	148	076
03051	KK	27/AUG/2014	27/AUG/2014	2	.U	040	.N	109	071
03052	TD	27/AUG/2014	27/AUG/2014	1	.U	065	.N	150	070
03057	UA	04/SEP/2014	04/SEP/2014	1	.U	067	.N	123	067
03060	PS	05/SEP/2014	05/SEP/2014	2	.U	075	.N	139	071
15019	TS	23/JAN/2014	23/JAN/2014	2	132	042	24.10	141	089
01019	DJ	28/FEB/2014	28/FEB/2014	2	135	067	36.76	146	071
06030	KK	24/MAR/2014	24/MAR/2014	2	138	043	22.57	160	098
17033	SS	01/MAY/2014	01/MAY/2014	2	140	036	18.36	118	081
11001	TC	25/JAN/2014	15/JAN/2014	1	140	040	20.40	136	060
20058	MS	15/MAY/2014	15/MAY/2014	2	140	046	23.46	120	080
13024	RS	23/APR/2014	23/APR/2014	2	140	049	25.00	117	068
03014	NC	18/FEB/2014	18/FEB/2014	2	140	058	29.59	130	080

Figure 6. Example of completeness metric of data

C—Consistent, ensuring data uniformity without contradictory information across the data collection forms/systems. Cross-checking related data to assure congruent values. The "consistency" can be examined by synchronizing different data sources and/or comparing data records from different datasets. For example, the researchers must investigate if a variable "sex" on a screening data collection form is "male" but on the enrollment form is "female." Query is needed when merging clinical database and laboratory worksheet, and data from the same specimen shows discordant information between reported values in clinical database versus those in the laboratory worksheet (Figure 7).

E—Enduring, securing data throughout the data lifecycle or the study duration. The data records should remain intact, accessible, and readable in a permanent and maintainable form within the study period for which they are intended.

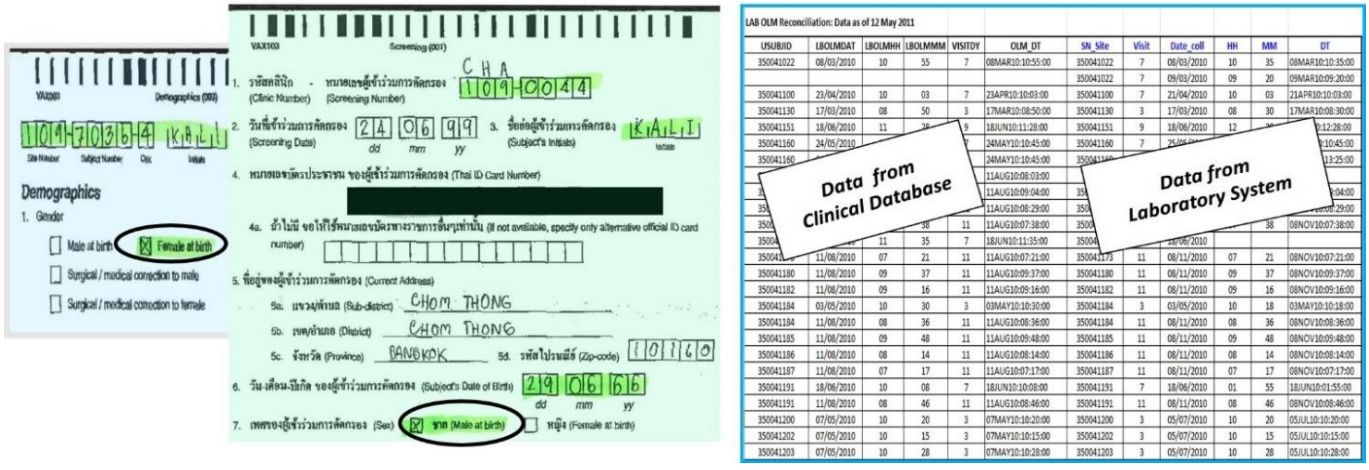

Figure 7. Examples of cross-checking of consistency between data

A—Available, warranting that data records are available throughout the data lifecycle or within the study period. The data records must be ready and accessible for review by responsible persons. The data records should be released or reported according to the pre-planned schedule.

Others

Other metrics that have been used in assessing the data integrity and data quality include uniqueness, validity, reliability, and relevance.

U—Uniqueness, verifying that data records are distinctively identifiable for each study participant. Each study participant should have only one study identification. There should be no duplicate data entries of the same event. For example, the data records of the same events/activities of a particular study participant must be deleted to avoid data repetitiveness. It is also important to cross-check the same study participant with different ID numbers to avoid multiple enrollments to the study (Figure 8).

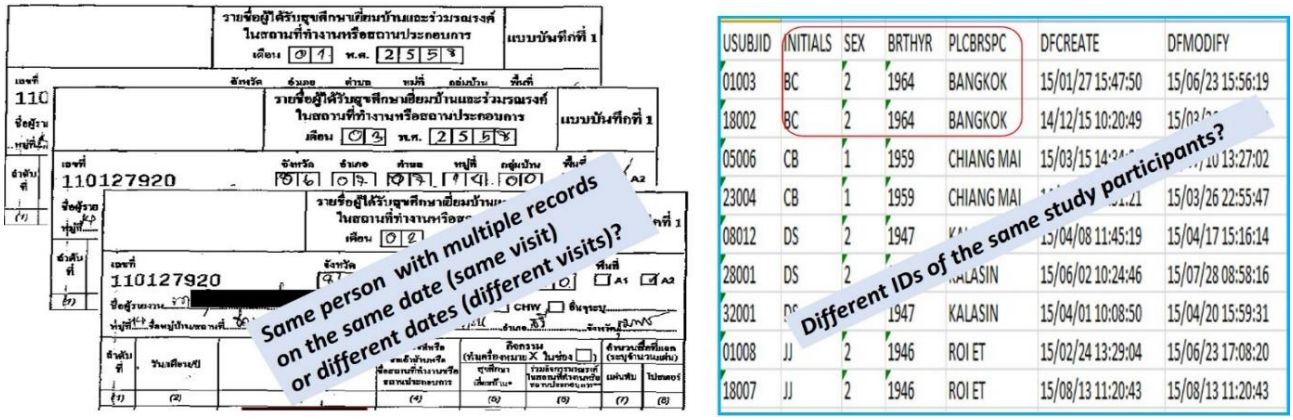

Figure 8. Examples of cross-checking of uniqueness in a set of data

V—Validity, validating the collected data whether they conform to acceptable format, type, or size according to the pre-set rules. The edit check program should be able to detect data values that are out-of-range or deviated from the normal range with unreasonable explanation. There should be standard coding for the open-ended variable. For example, the system should have edit check for unusual white blood cell count. The verbatim of adverse event as reported by the hospital staff must be converted to standard coding scheme (e.g., ICD-11 or MedDRA coding) for data analysis. It is important to train research staff to enter the data according to the data collection manual, e.g., not reporting drug name for the variable that should capture adverse event data.

Figure 9. Examples of checking validity of data

R—Reliability, assuring that analysis of data produces consistent results over time within an individual study participant and/or across different records within the dataset. Reliability of the data can be detected by the inconsistency and/or illogical reason of the data values. Such quality of the data may be observed by basic calculation or after performing data analysis. For example, is it possible that a study participant who has been reported with "confirmed HIV positive" for several visits became "not infected" in the last visit? Is it correct that survival time of the patients, calculated from (date last visit–date diagnosis), are negative, extremely high, or zero? (Figure 10).

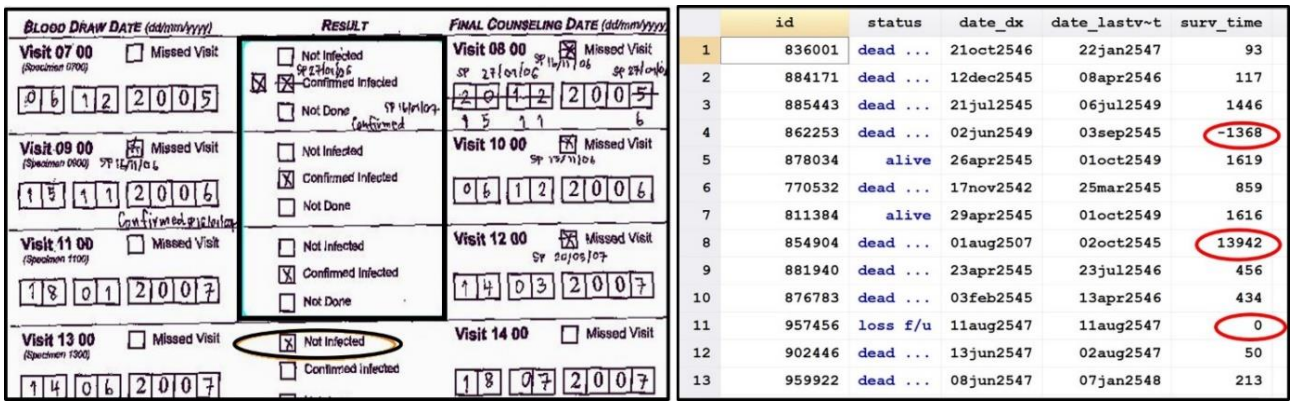

Figure 10. Examples of checking reliability of data within the dataset

R—Relevance, affirming that the data collected according to the protocol requirement. This metric reflects back to the step of data design—what critical variables, including type and size, are needed to answer the research question. Edit check program for the data entry and validation plan are helpful to ensure the collected data are within the format and scope of the study. For example, data values are set according to the validation plan by automated edit check program within the data entry system (Figure 11).

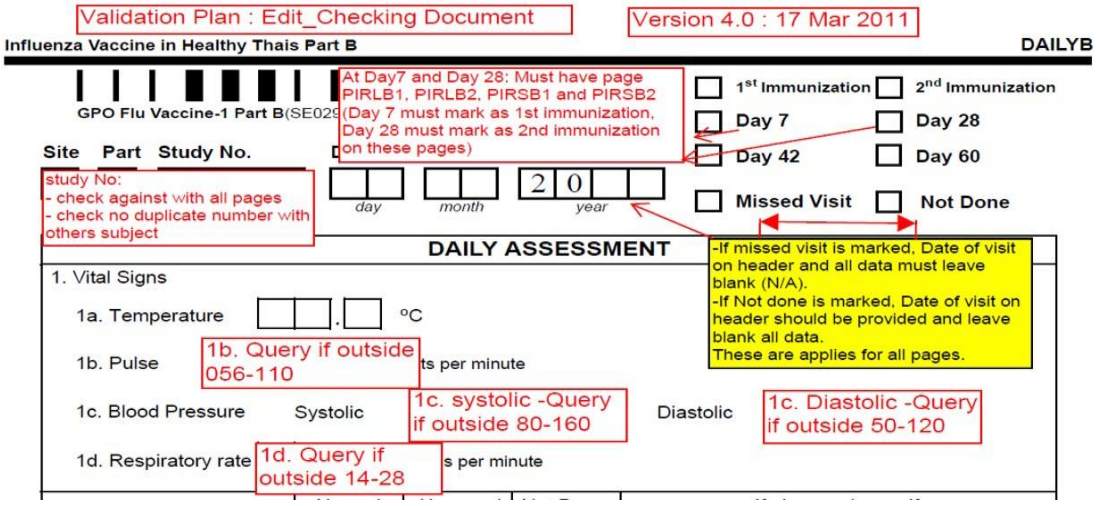

Figure 11. Example of checking data values on data entry screen

Framework for Monitoring Data Quality

One of the frameworks proposed in literature for monitoring data quality include, but not limited to, the followings: ratio of data to errors (how many issues are raised?), number of empty values (how many empty fields are there?), data transformation error rate (if the data are transformed, how often that they are performed incorrectly?), data storage or management costs (how much is the cost of data archival or maintenance?).[21]

In assuring data quality, the data management procedures are required to leave the so-called "audit trail" which will show traceable activities from initial data entry to interim and final reports.[1,20,23] The aim of audit trail is to confirm the whole process such that: the data reported are the data analyzed; the data analyzed are the data recorded on data collection tools; the data on the data collection tools are the data generated from original source; and the data generated are compliant to the study protocol.[23]

Conclusion

A good data quality management with help improve the trustworthiness of your data.[22] Trustworthiness is a product of the people, systems and processes that enable and support the management and production of data.[5] It is important to train research team on data management and governance best practices and provide ongoing monitoring and reeducation.[18] It is essential to assure the quality of study conduct and the trustworthiness of data to achieve the reliable study results.

Suggested Citation

Kaewkungwal J. The grammar of science: it's all about trustworthiness of your data. OSIR. 2024 Sep;17(3): 165–72. doi:10.59096/osir.v17i3.270958.

References

1. International council for harmonization of technical requirements for pharmaceuticals for human use: good clinical practice E6(R3). [cited 2024, August 15]. 73 p. <https://database.ich.org/sites/default/files/ICH_E6%28R3%29_DraftGuideline_2023_0519.pdf>
2. World Health Organization. WHO expert committee on specifications for pharmaceutical preparations: fifty-fifth report [Internet]. Geneva: World Health Organization; 2021. Annex 4: guideline on data integrity; [cited 2024 Aug 15]; p 135–159. <https://cdn.who.int/media/docs/default-source/medicines/norms-and-standards/guidelines/inspections/trs1033-annex4-guideline-on-data-integrity.pdf?>
3. U.S. Food and Drug Administration. Data integrity and compliance with drug CGMP: questions and answers [Internet]. Silver Spring (MD): U.S. Food and Drug Administration; 2018 Dec [cited 2024 Aug 15] <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/data-integrity-and-compliance-drug-cgmp-questions-and-answers>
4. European Medicines Agency. Guidance on good manufacturing practice and good distribution

practice: questions and answers [Internet]. Amsterdam: European Medicines Agency; [cited 2024 Aug 15] <https://www.ema.europa.eu/en/human-regulatory-overview/research-development/compliance-research-development/good-manufacturing-practice/guidance-good-manufacturing-practice-good-distribution-practice-questions-answers#data-integrity-7108>

5. Office for Statistics Regulation, UK Statistics Authority. Code of Practice for statistics: Trustworthiness [Internet]. London: UK Statistics Authority; [cited 2024 Aug 15]. <https://code.statisticsauthority.gov.uk/the-code/trustworthiness/>
6. Atlan. Data governance vs. data management: what's the difference? [Internet]. Singapore: Atlan Technologies; [updated 2023 Apr 3; cited 2024 Aug 15]. <https://atlan.com/data-governance-vs-data-management/>
7. Atlan. Data quality in data governance: the crucial link that ensures data accuracy and integrity [Internet]. Singapore: Atlan Technologies; [updated 2023 Jul 24; cited 2024 Aug 15]. <https://atlan.com/data-quality-in-data-governance/>
8. DQLabs. Challenges and best practices of data cleansing [Internet]. Pasadena (CA): DQLabs; [cited 2024 Aug 15]. <https://www.dqlabs.ai/blog/challenges-and-best-practices-of-data-cleansing/>
9. Bothra R. Data quality management techniques and best practices [Internet]. San Francisco: Hevo Data Inc; 2024 May 3 [cited 2024 Aug 15]. <https://hevodata.com/learn/what-is-data-quality-management/>
10. Collibra. Data quality and data governance: where to begin? [Internet]. New York: Collibra; [updated 2023 Dec 14; cited 2024 Aug 15]. <https://www.collibra.com/us/en/blog/data-quality-vs-data-governance>
11. Tableau. Data management vs. data governance: the difference explained [Internet]. Seattle: Tableau Software; [cited 2024 Aug 15]. <https://www.tableau.com/learn/articles/data-management-vs-data-governan>
12. Mucci T, Stryker C. What is data integrity? [Internet]. New York: IBM; 2024 Apr 5 [cited 2024 Aug 15]. <https://www.ibm.com/topics/data-integrity>
13. Loshin D. The practitioner's guide to data quality improvement [Internet]. Philadelphia: Elsevier Inc; 2011 [cited 2024 Aug 15]. <https://www.sciencedirect.com/book/9780123737175/the-practitioners-guide-to-data-quality-improvement>
14. Pharmaceutical Consultancy Services. The difference between ALCOA and ALCOA+ (plus) [Internet]. Woerden: Pharmaceutical consultancy, training and audits; 2023 Oct 30 [cited 2024 Aug 15]. <https://www.pcs-nl.com/en/post/difference-between-alcoa-and-alcoa-plus>
15. Novogroder I. Data quality management: tools, pillars, and best practices. Campbell (CA): lakeFS; 2023 Nov 30 [cited 2024 Aug 15]. <https://lakefs.io/data-quality/data-quality-management/>
16. Cantor K. What Is the difference between data cleaning and data cleansing? [Internet]. Indianapolis: P3Adaptive; 2024 May 27 [cited 2024 Aug 15]. <https://p3adaptive.com/what-is-the-difference-between-data-cleaning-and-data-cleansing/>
17. DQOps. Data cleaning vs data cleansing [Internet]. Warsaw: DQOps; [updated 2024 Jun 24; cited 2024 Aug 15]. <https://dqops.com/data-cleaning-vs-data-cleansing/>
18. Acceldata. What is data quality management? Campbell (CA): Acceldata; 2024 May 3 [cited 2024 Aug 15]. <https://www.acceldata.io/article/what-is-data-quality-management>
19. Qualifyze. What is ALCOA data integrity? [Internet]. Frankfurt: Qualifyze; 2022 Oct 11 [cited 2024 Aug 15]. <https://www.qualifyze.com/resources/blog/what-is-alcoa-data-integrity/>
20. Society for Clinical Data Management. Good clinical data management practices (2024 edition). Brussels: Society for Clinical Data Management; [cited 2024 Aug 15]. <https://scdm.org/gcdmp/>
21. Actian. What is data quality management? [Internet]. Santa Clara (CA): Actian Corporation; [cited 2024 Aug 15]. <https://www.actian.com/what-is-data-quality-management/>
22. Bauman J. Data quality management: what you need to know [Internet]. Cary: SAS Institute Inc; [cited 2024 Aug 15]. <https://www.sas.com/en_th/insights/articles/data-management/data-quality-management-what-you-need-to-know.html>
23. Rondel RK, Varley SA, Webb CF. Clinical data management. 2nd ed. New York: John Wiley & Sons, Ltd: 2000. 368 p.