Data Ordinality: An Important Concept in Reducing Medical and Research Errors

This manuscript (permalink) was automatically generated from fmaguire/data_ordinality_paper@90f38c3 on October 15, 2019.

Authors

- **Finlay Maguire**
  0000-0002-1203-9514 · fmaguire · fmaguire
  Faculty of Computer Science, Dalhousie University · Funded by Donald Hill Family Fellowship in Computer Science

- **Michael A. Graven**

  Department of Paediatrics, Faculty of Medicine, Dalhousie University

- **Noni E. MacDonald**

  Department of Paediatrics, Faculty of Medicine, Dalhousie University

Data Ordinality

Adverse effects of medical treatment are responsible for 2.8% of all deaths in the US [1]. A major cause of these adverse medical outcomes is error in the transmission and communication of information between healthcare workers (HCW) [2]. These errors are very common, with manually entered pathology data having errors in 76% of patients records [3], and transcription errors (misinterpreation of written/spoken orders and mistakes copying prescriptions) responsible for 2-14% of medication errors [4]. Additionally, these errors are not evenly distributed across medical information with some specific pieces of data having error rates as high as 26.9% [5].

While errors in patient data can have disastorous consequences for that patient, these mistakes can also have much more wide-ranging impact. Retrospective and prospective analysis of datasets derived from patient chart data make up a huge proportion of clinical research publications [6]. These datasets are often used to construct or update research databases that are re-used by large numbers of later studies. Some of these databases may have error rates as high as 26.9% for some pieces of data [5]. Therefore, the errors in this clinical data (whether introduced in the underlying patient data or in the later extraction and cleaning) can and likely does lead to incorrect erroneous study outcomes [7]. One recent high-profile example of this involved errors in the large and well-respected UK Biobank database [8]. These distorted study results have the potential to harm huge numbers of patients by contributing to incorrect changes in the standards of care.

It is generally cheaper and easier to prevent the introduction of errors into data than to retrospectively attempt to correct the data. If we want to try and prevent these adverse medical and experimental outcomes we need to minimise the number of opportunites for error to be introduced. Humans, even well trained ones, are fallible. This means that the more individuals a piece of data passes through, the more likely it is to acquire errors. An intuitive example of this effect is that of the childhood “whisper game”. In this, a message is whispered from person to person down a line of people. Upon the last person in the line speaking the message aloud the message is typically drastically different from the original message. Each person listening and repeating this message allows the original information to become increasingly distorted and warped. Medical information, which requires shared between large networks of HCWs over extended periods of times is thus particularly prone to these distortions. Therefore, minimising the number of error-prone human links in these chains is the best way to mitigate the errors.

To help researchers and clinicians understand and conceptualise this we present the principle of “data ordinality”. Ordinality is a term used in mathematics to denote the order of objects within an ordered sequence, e.g., 1st, 2nd, 3rd. We can apply this idea to data, by using it describe how many times a piece of data has been manually recorded or entered into a system. For example, 1st order data would be data that has been directly extracted or printed out from a device such as a digital thermometer.
When this temperate is written down by a HCW into a patient’s chart it would now be 2nd order data. Finally, when a researcher extracts this temperature from the chart it is now 3rd order data. In our intuitive example, data ordinality would be the number of times the message is whispered between individuals in the game of telephone/whisper. Errors increase as data ordinality increases, much as the distortion of the message becomes greater the more people are playing the whisper game.

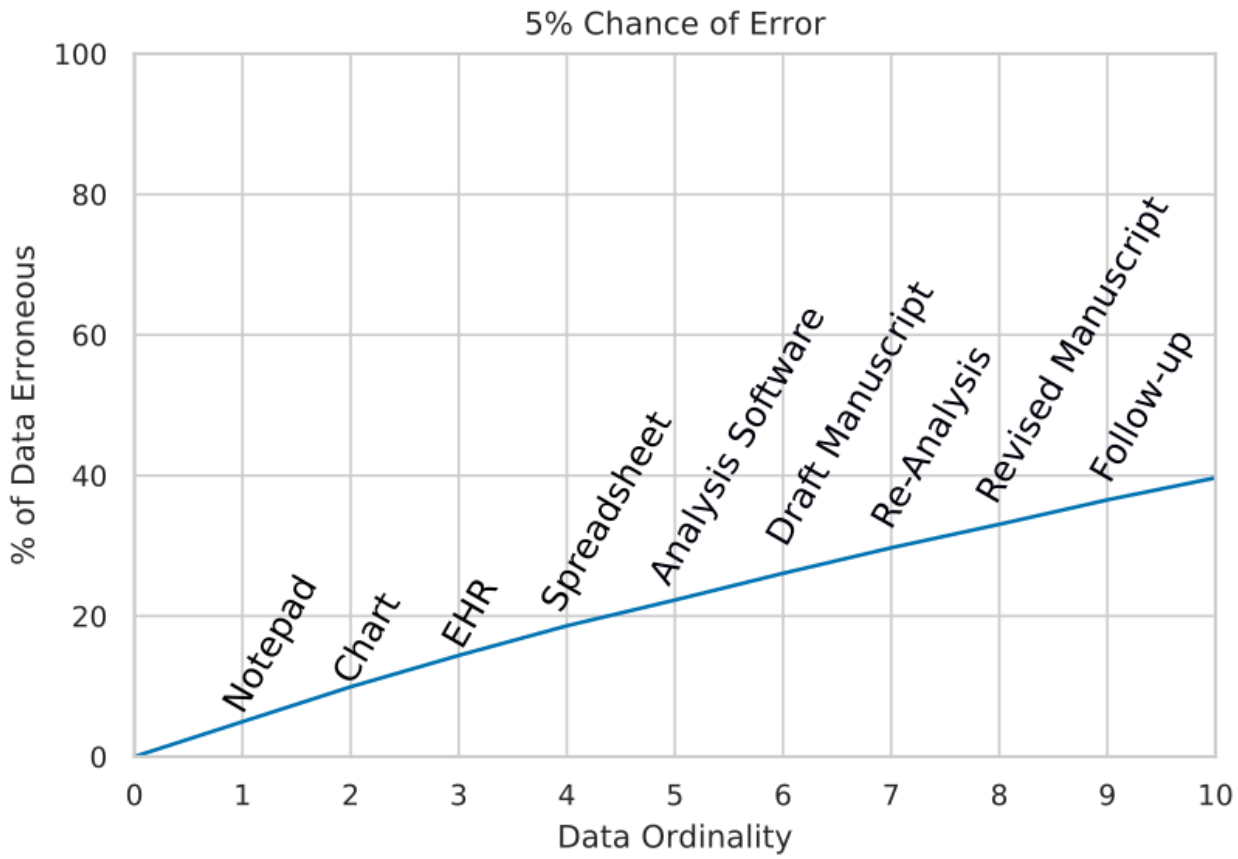

Figure 1: Propagation of Error with Data Ordinality. Assuming an arbitrary uniform 5% chance of error in the copying of each data-point at each stage of copying. In reality different pieces of data and different stages will have drastically different error rates.

An effective way to reduce data ordinality is the adoption of clinical information technology (IT) such as electronic health records, electronic prescription systems, and so on. The adoption of systems like this have been shown to reduce hospital-related morbidity, mortality, and costs [9]. With immediate input of data into these systems, ideally directly from the medical equipment, we can greatly reduce data ordinality because any subsequent user of the information can access it on the database. No matter how many people need access to the data it will not require being repeatedly transcribed as it would for a piece of data only recorded on paper.

It is, however, not sufficient to just enter the information into any digital form as without planning the ordinality of digital data can still increase. For example, a dataset could be entered into a spreadsheet, shared with one researcher. That researcher could make some accidental changes due to either user error or computer glitches (partial data transfer, bugs, automated reformatting). This mangled data is then sent to another collaborator who introduces their own errors before finally being uploaded to a publication database having accumulated ordinality and errors. A pervasive example of this in the life sciences is the mangling of gene names through automated formatting of them as dates by Excel [11]. These mangled names are then uploaded to central databases where they continue to proliferate as other researchers make use of them. This means, whether it is a physical or an IT system, it is vital to have a single authorative source that can only be changed in a restricted audited way. These datasets and any updates to them should be tracked using a version control system and be thoroughly backed-up in order to ensure data integrity. Indeed, this kind of data integrity and versioning requirements are enumerated in many national laws and regulations governing clinical trials and pharamceutical manufacture (e.g. US Federal Drug Association's 21CFR11).

While most researchers, hospital administrators, and clinicians will understand the idea of data ordinality, we believe explicitly codifying the idea will be useful in many areas of pratice and system

design. Errors are invetible but by thinking about data ordinality, the utility of interventions that generate authoritative carefully controlled and preserved information sources becomes clear. A cconcept of data ordinality also makes it easier to identify dangerous practices that lead to the medical “whisper game” and so helps to mitigate these situations. Without a clear explicit conception of data ordinality it becomes much harder to build these ideas into the design of clinical systems and research protocols. The more difficult it is to design systems that take this into account, the more likely it is produce a system that causes data errors. Ultimately, we belive that the concept of data ordinality is a very useful mental framework that will help clinicians be more mindful of data in both practice and research.

References

1. **Association of Adverse Effects of Medical Treatment With Mortality in the United States**
Jacob E. Sunshine, Nicholas Meo, Nicholas J. Kassebaum, Michael L. Collison, Ali H. Mokdad, Mohsen Naghavi
JAMA Network Open (2019-01-18) https://doi.org/gf9gds
DOI: 10.1001/jamanetworkopen.2018.7041 · PMID: 30657530 · PMCID: PMC6484545

2. **Debunking the myth that the majority of medical errors are attributed to communication**
Timothy C Clapper, Kevin Ching
Medical Education (2019-09-11) https://doi.org/gf9vtc
DOI: 10.1111/medu.13821 · PMID: 31509277

3. **Error rates in a clinical data repository: lessons from the transition to electronic data transfer–a descriptive study.**
Matthew KH Hong, Henry HI Yao, John S Pedersen, Justin S Peters, Anthony J Costello, Declan G Murphy, Christopher M Hovens, Niall M Corcoran
BMJ open (2013-05-28) https://www.ncbi.nlm.nih.gov/pubmed/23793682
DOI: 10.1136/bmjopen-2012-002406 · PMID: 23793682 · PMCID: PMC3657671

4. **Drug-Related Problems in Hospitals**
Anita Kr??henb??hl-Melcher, Raymond Schlienger, Markus Lampert, Manuel Haschke, J??rgen Drewe, Stephan Kr??henb??hl
Drug Safety (2007) https://doi.org/bg3z3z
DOI: 10.2165/00002018-200730050-00003 · PMID: 17472418

5. **Analysis of data errors in clinical research databases.**
Saveli I Goldberg, Andrzej Niemierko, Alexander Turchin
AMIA ... Annual Symposium proceedings. AMIA Symposium (2008-11-06)
https://www.ncbi.nlm.nih.gov/pubmed/18998889
PMID: 18998889 · PMCID: PMC2656002

6. **Chart reviews in emergency medicine research: Where are the methods?**
EH Gilbert, SR Lowenstein, J Koziol-McLain, DC Barta, J Steiner
Annals of emergency medicine (1996-03) https://www.ncbi.nlm.nih.gov/pubmed/8599488
DOI: 10.1016/s0196-0644(96)70264-0 · PMID: 8599488

7. **Modelling of errors in databases.**
Steve Gallivan, Christina Pagel
Health care management science (2008-03) https://www.ncbi.nlm.nih.gov/pubmed/18390166
PMID: 18390166

8. **Retraction Note: CCR5-Δ32 is deleterious in the homozygous state in humans**
Xinzhu Wei, Rasmus Nielsen
Nature Medicine (2019-10-08) https://doi.org/gf9vtb
DOI: 10.1038/s41591-019-0637-6 · PMID: 31595084

9. **Clinical Information Technologies and Inpatient Outcomes**
Ruben Amarasingham, Laura Plantinga, Marie Diener-West, Darrell J. Gaskin, Neil R. Powe
Archives of Internal Medicine (2009-01-26) https://doi.org/bzmhbp
DOI: 10.1001/archinternmed.2008.520 · PMID: 19171805

10. **Does the Leapfrog program help identify high-quality hospitals?**
Ashish K Jha, E John Orav, Abigail B Ridgway, Jie Zheng, Arnold M Epstein
Joint Commission journal on quality and patient safety (2008-06)
https://www.ncbi.nlm.nih.gov/pubmed/18595377
PMID: 18595377

11. **Gene name errors are widespread in the scientific literature**
Mark Ziemann, Yotam Eren, Assam El-Osta
Genome Biology (2016-08-23) https://doi.org/bqt3
DOI: 10.1186/s13059-016-1044-7 · PMID: 27552985 · PMCID: PMC4994289

12.
Barry R Zeeberg, Joseph Riss, David W Kane, Kimberly J Bussey, Edward Uchio, W Marston Linehan, J Carl Barrett, John N Weinstein
BMC Bioinformatics (2004) https://doi.org/fcvtjq
DOI: 10.1186/1471-2105-5-80 · PMID: 15214961 · PMCID: PMC459209