Additive Theory of Error Generation and Correction Derived from & Applied to Clinical Research Data Management Modeling Data Accuracy in Clinical Research Data Management

ABSTRACT----------

A quantitative model for data quality planning in clinical research data management does not exist. Thus, data collection and management processes in clinical research are currently designed according to practice, intuition and individual experience, and subsequently formalized in organizational quality systems. Inspired by, but different from Orr's 1998 System Theory and data quality work, we employ a control theory approach to model data accuracy through a series of data processing steps. Expressions for the interim and outgoing data accuracy from a data processing process are derived from first principles. The model is tested at known boundaries, benchmarked with two previous models, and benchmarked with error generation and correction rates consistent with those in the clinical research data quality literature. This first generation model enables prospective evaluation of candidate process paths and methodology with respect to data accuracy. As such, the model is beneficial to practitioners.

BIOGRAPHY----------

Meredith Nahm, MS
Associate Director for Clinical Research Informatics,
Duke Translational Medicine Institute

A member of the Duke community for ten years, Ms. Nahm previously served as the Director of Clinical Data Integration at the Duke Clinical Research Institute. She has over 15 years of experience in research data management. She has served on the boards of the Society for Clinical Data Management (SCDM) and the Clinical Data Integration Standards Consortium (CDISC) and currently co-chair's the Clinical Interoperability Council in Health Level 7.

Ms. Nahm authored the Measuring and Assuring Data Quality chapters in the Good Clinical Data Management Practices Document published by the Society for Clinical Data Management, and has authored several papers on clinical research data quality. Ms. Nahm is currently pursuing a PhD in Biomedical informatics at the School of Health Information Sciences, University of Texas at Houston. Her research interests include data quality, knowledge representation, and clinical research informatics.

Leonard White, MS, PE
Senior Electrical Engineer
Stanford White, Inc.

Leonard W. White received his MS in Electrical and Computer Engineering from NC State University. He is one of the founding partners and former Senior Principal of Stanford White, Inc., a mid-sized engineering firm specializing in engineering services for the construction industry. Mr. White is a registered professional engineer in eight states, a Registered Communications Distribution Designer (RCDD), serves on the NFPA-99 hospital electrical systems committee and is a senior member of IEEE. His area of specialization is power quality. He is presently pursuing a PhD in Electrical Engineering at NC State University.

Constance M. Johnson, RN, PhD
Assistant Professor
Duke University School of Nursing

Dr. Johnson directs the Informatics program at the Duke University School of Nursing. An informatician with interdisciplinary training in nursing and health informatics, Dr. Johnson has more than 20 years of experience in research and informatics in the areas of health promotion and disease prevention. In addition to developing and directing the development of numerous large databases, as well as user interfaces in the areas of obstetrics/neonatology, cancer prevention, and cancer genetics, Dr. Johnson has extensive experience with large population studies. She has done research in preterm labor prevention, health care informatics, mental models, human-centered interface and web design, colorectal cancer prevention, information visualization, and cancer risk models. While at the University of Texas Health Science Center, Dr. Johnson studied under an F38-Fellowship from the National Library of Medicine. She has given numerous national peer-reviewed conference presentations and has been an author on numerous articles.

Todd R. Johnson, PhD
Associate Dean for Academic Affairs
School of Health Information Science
University of Texas at Houston

Dr. Johnson received his PhD from Ohio State University. His research applies cognitive science, computer science, and human factors engineering to solve informatics problems. Medical Device Usability and Safety, focusing on Human-centered interface design for Patient Safety and Quality, Decision support, Computer models of human problem-solving behavior and learning, Ontologies and knowledge sharing
Dr. Johnson is an expert in cognitive science in healthcare, an area that improves healthcare and biomedical decision making by designing processes, software, and devices that match the needs and cognitive capabilities of those who use them. His current work focuses on two areas: 1) Improving patient safety by reducing medical errors caused by poor device and software interfaces, as well as errors that arise due to pressures placed on caregivers by the healthcare

system in which they work; and 2) Improving decision making and efficiency through user-centered software design and decision support systems.

Jiajie Zhang, PhD
Associate Dean for Research
School of Health Information Science
University of Texas at Houston

Dr. Zhang is a cognitive scientist with interdisciplinary training in cognitive psychology, computer science, and neurosciences. He has done research in biomedical informatics, cognitive science, human-centered computing, user interface design, information visualization, medical error, decision making, and computational cognitive modeling. He has authored numerous articles, book chapters, and peer-reviewed proceedings papers. He has been the principal investigator or co-investigator on many grants from NASA, Office of Naval Research, Army, NIH, James S. McDonnell Foundation, and other funding agencies. He has given numerous conference presentations and invited presentations at other institutions, and organized and participated in many symposia and panels at international and national conferences. He has also served on several NIH review panels. Dr. Zhang was a recipient of John P. McGovern Outstanding Teacher Award in 2002. Dr. Zhang is an elected Fellow of American College of Medical Informatics.

Modeling Data Accuracy in Clinical Research Data Management

Meredith Nahm, MS[a,b], Leonard W. White, MS, PE[c], Constance Johnson, PhD[d], Todd Johnson, PhD[b], Jiajie Zhang, PhD[b]

a Duke Translational Medicine Institute, Duke University, Durham NC,
b University of Texas at Houston School of Health Information Sciences, Houston Tx,
c Stanford White, Inc., Raleigh NC
d Duke University School of Nursing, Durham, NC

Information Quality Industry Symposium (IQIS)
July 15-17, 2009
Cambridge Massachusetts

Duke Translational Medicine Institute
DUKE UNIVERSITY MEDICAL CENTER

The University of Texas
School of Health Information Sciences at Houston

Topics

- Background: Clinical research data management
- Data quality in clinical research data mgt.
- Two related models
- New model
  - Axioms
  - Theory
  - Benchmarks and Application
  - Strengths and Limitations
- Conclusions

Problem Statement

- **"Mass customization":** data processes and methodology are constructed for each clinical trial
  - Scientific differences → data differences
  - Outsourcing → Fragmentation of research programs
  - Methodological non-uniformity

    Unsynthesized evidence base → individual decisions, apprentices
- **A predictive of model data accuracy obtainable from candidate processes would be helpful**

Quote, Karen Koh, Personal communication, 2009

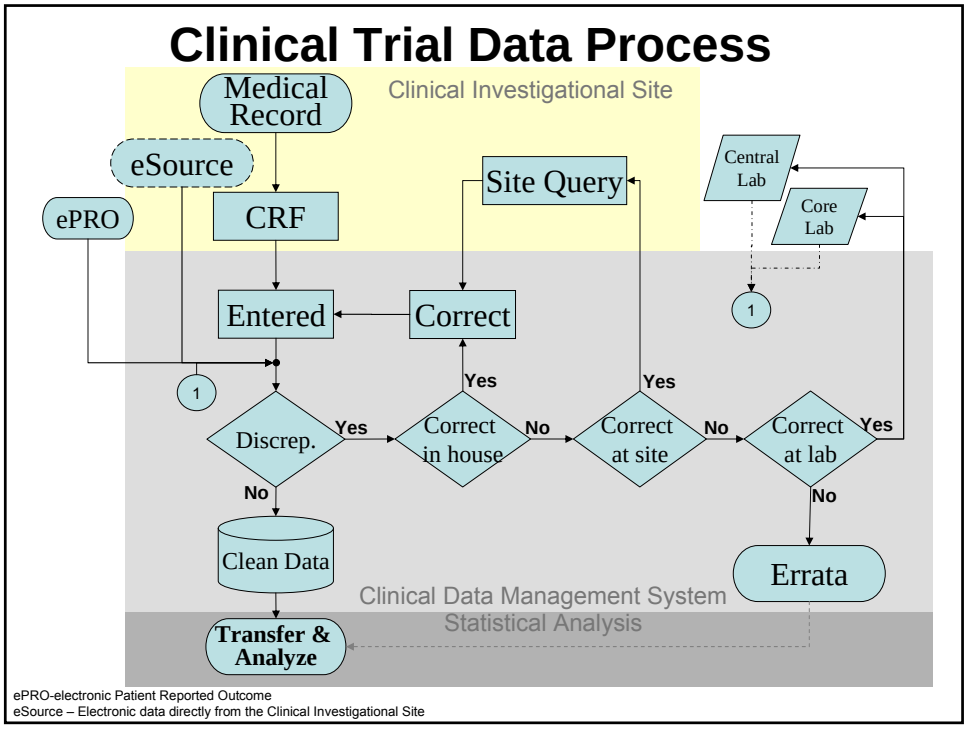
Clinical Research R&D Spending

- Canada 2007, $1.3 billion
  - 47% of which is spent on clinical trials
- US NIH 2008, $28.9 billion
- US Pharma Industry 2007 $44.5 billion PHRMA members
  - 67% of which spent on clinical trials

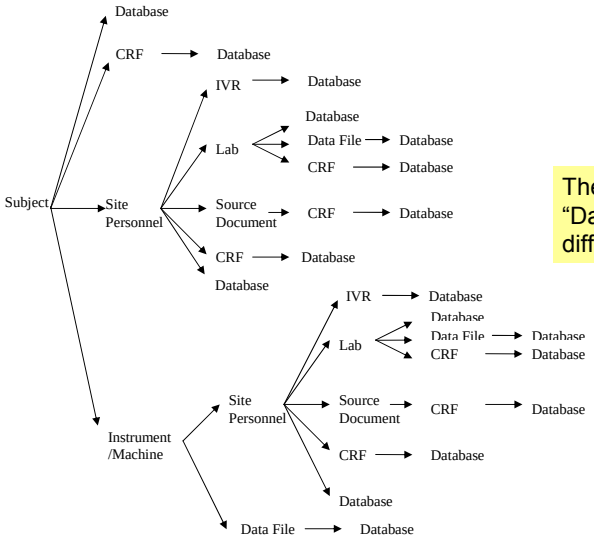
"The pharmaceutical industry is, one of the most research intensive industries in the United States. Pharmaceutical firms invest as much as **five times more in research and development,** relative to their sales, than the average U.S. manufacturing firm." — Congressional Budget Office October 2006

Patented Medicine Prices Review Board, Annual Report, 2007. Analysis of research and development expenditure, available from http://www.pmprb-cepmb.gc.ca/english/view.asp?x=1068&mid=864 accessed on November 30, 2008.

Pharmaceutical Research and Manufacturers of America, *Pharmaceutical Industry Profile 2008* (Washington, DC: PhRMA, March 2008).



Data Paths in Clinical Research



They all end with "Database", but after differing numbers of steps.

Clinical Research Data Management

Activities
- Data collection
- Data processing (entry, cleaning, coding)
- Data integration
- System design, testing, and support

Data from
- Patients
- Electronic medical records
- Electronic devices
- Paper forms …

Data types
- Text & numbers
- Images
- Signals
- Biological samples

Fertile environment for the study of data quality issues.

Data Processing Methods in Clinical Research

- Entry
  - Optical
  - Key entry (single, double, double variations)
  - Patient handheld device
  - Voice, …
- Cleaning
  - On-screen error traps
  - Batch programming
  - Manual review / visual verification
  - Aggregate data checks
- Coding
  - Manual (many variations)
  - Autoencoding
- …

Are there particularly good or bad combinations ?

Current State Clinical Research Data Management

- A quantitative framework for quality planning in clinical research data management does not exist.
- Data collection and management processes currently designed according to practice, intuition and individual experience, then
- Formalized in organizational quality system through policies and standard operating procedures (SOPs) GCDMP, Assuring Data Quality section

Society for Clinical Data Management, 2007. Good Clinical Data Management Practices Document. Available from www.scdm.org

Data Processing as a System

- As suggested by Orr's System Theory conceptualization, data processes can be represented as a system.
- Inspired by but different from Orr's work,
- We employ a control theory approach
  - derive from first principles, expressions for the interim and outgoing data accuracy from a data processing process

ORR, K. 1998. Data Quality and Systems Theory. *Communications of the ACM 41*, 6.

Gardiner Model

Gardiner, 1978 applied basic probability model to prediction of error rates. This model accounted for error generation but did not account for error correction.

Assumed independent events.
Assumed each event has **same** independent error probability, p.
Probability of no error of any one event is (1-p)
N is the number of events.
By the multiplication rule (**joint, intersection probability**), probability of N independent events is
$(1-p)_1*(1-p)_2*(1-p)_3 \ldots (1-p)_N$
We want the compliment, i.e. error rate, so $1 - (1-p)_1*(1-p)_2*(1-p)_3 \ldots (1-p)_N$

Error rate after N steps
Where error occurs in all steps $= 1 - (1 - p)^{\mathbf{N}}$

Even low probabilities at each step result in substantial error rate of the whole.
- represents error rate where error occurs in all steps
- another option: addition rule, union, (A or B or both, inclusive or)
- another option: exclusive or (either but not both), $p_1 + p_2 + p_3$ most applicable

Gardiner, Richard C. (1978) Quality Considerations in medical Records Abstracting Systems. Journal of Medical Systems, 2:1; 31-43

Ma Model

- Probability modeling approach with record as measurement unit
- Process based approach, for three process variations below

Incoming error from data entry

Error correction

Outgoing error

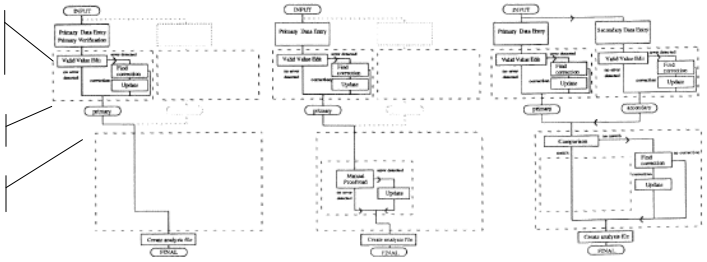


Figure 1a. Primary VVE Only *Figure 1b. Primary VVE and Manual Proofread* *Figure 1c. Intelligent Entry and Redundancy*

$z_1 = e_3 = (1-kd_1)e_1$

$z_2 = r_{01}e_3 + (1-k)[r_{10}(1-e_3) + r_{11}e_3]$

$z_3 = e_3 - ck(e_3 - me_3e_4)$

Random variable Z is a Binomial (N, z) defined as the sum of N i.i.d. Bernoulli random variables
Z = 1 means record in error in final data set
e is an error rate,
d is an error detection rate of valid value edit checks, c is conditional probability of detected given that error exists and proofread
k is a correction rate conditional on an update needed,
r is a error detection rate from manual proof reading

Ma, Mei-mei Juliana (1986) A modeling Approach to System Evaluation in Research Data Management. Dissertation, University of North Carolina, Chapel Hill

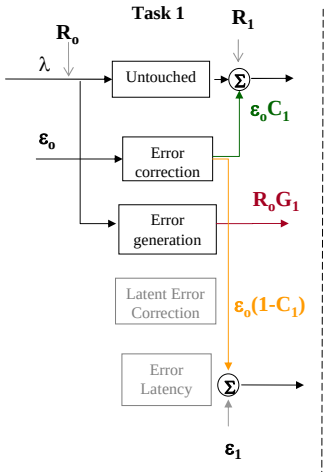
Similarities

- Recognition that each process step has potential to generate error
- A value is either correct or incorrect
- Error generation and correction rates
- Random as opposed to systematic errors (Ma)
- Independence is assumed (Ma to extent possible)

Our Model: Axioms

- Data are neither created nor destroyed.
- Data are either accurate or inaccurate.
- Each data value travels through a specified path; such paths may branch.
- At any data processing step, an **accurate** value may be untouched, or rendered incorrect.
- Likewise, at any data processing step, an **inaccurate** value may be untouched, or rendered incorrect.

Summing the Error Generation and Correction for the First Processing Step



$$R_1 = R_o - R_o G_1 + \varepsilon_o C_1$$
$$\varepsilon_1 = \varepsilon_o (1 - C_1)$$

R_o and ε_o are the input or initial values.

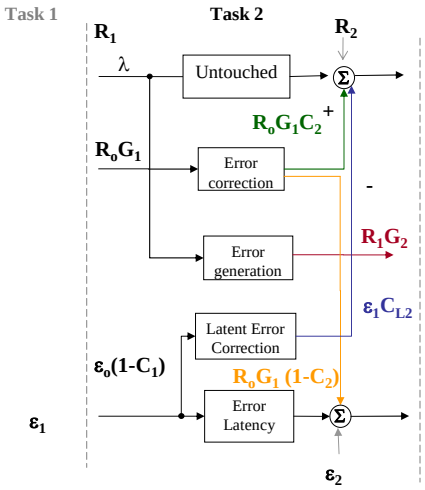
Each task may touch all or a fraction, λ, of the data.

R_i represents the number of accurate data values outgoing from the i^{th} process step.

ε_i, is the number of values in error outgoing from the i^{th} process step.

G and C are the error generation and correction rates

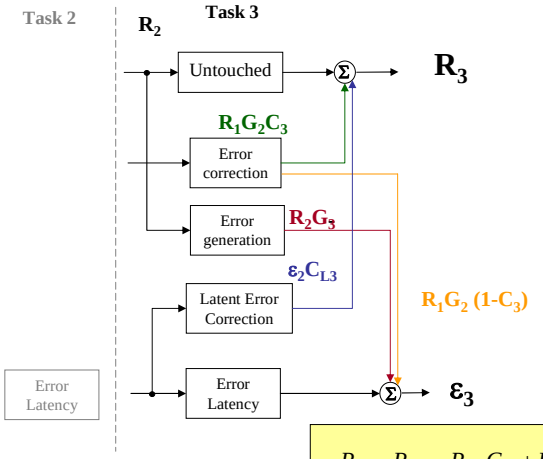
Summing the Error Generation and Correction for the i^{th} Processing Step



$$R_i = R_{i-1} - R_{i-1} G_i + R_{i-2} G_{i-1} C_i + \varepsilon_{i-1} C_{iL}$$
$$\varepsilon_i = \varepsilon_{i-1} - \varepsilon_{i-1} C_{iL} + R_{i-2} G_{i-1} (1 - C_i)$$

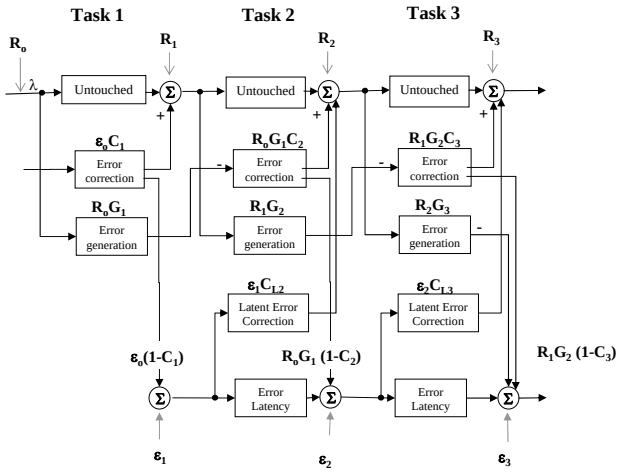
Error Latency

Outgoing Accuracy & Error

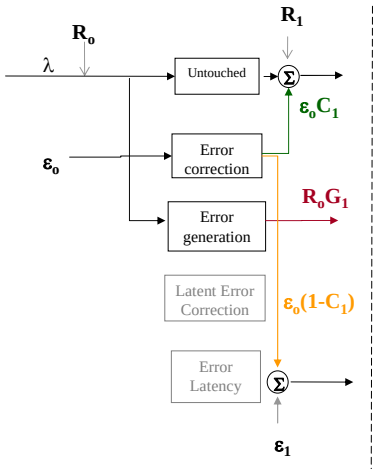


$$R_n = R_{n-1} - R_{n-1}G_n + R_{n-2}G_{n-1}C_n + \varepsilon_{n-1}C_{nL}$$
$$\varepsilon_n = \varepsilon_{n-1} - \varepsilon_{n-1}C_{nL} + R_{n-2}G_{n-1}(1-C_n) + R_{n-1}G_n$$

Putting it all Together



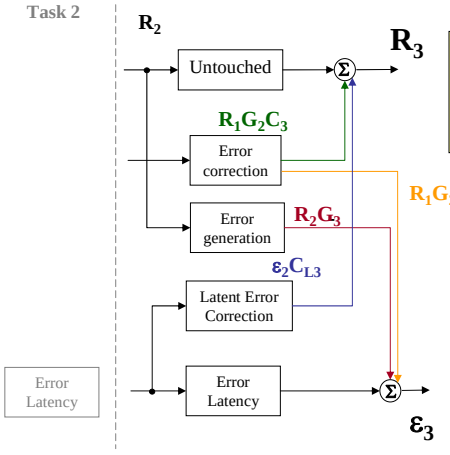
Testing at the Boundaries, R_1 and ε_1



$$R_1 = R_o - R_o G_1 + \varepsilon_o C_1$$
$$\varepsilon_1 = \varepsilon_o (1 - C_1)$$

Error correction C	Error generation G: $G \to 0$	Error generation G: $G \to 1$
$C \to 0$	$R_1 \to R_o$, $\varepsilon_1 \to \varepsilon_o$	$R_1 \to 0$, $\varepsilon_1 \to \varepsilon_o$
$C \to 1$	$R_1 \to R_o + \varepsilon_o$, $\varepsilon_1 \to 0$	$R_1 \to \varepsilon_o$, $\varepsilon_1 \to 0$

Testing at Outgoing Boundaries



$$R_n = R_{n-1} - R_{n-1} G_n + R_{n-2} G_{n-1} C_n + \varepsilon_{n-1} C_{nL}$$
$$\varepsilon_n = \varepsilon_{n-1} - \varepsilon_{n-1} C_{nL} + R_{n-2} G_{n-1} (1 - C_n) + R_{n-1} G_n$$

Error correction C	Error generation G: $G \to 0$	Error generation G: $G \to 1$
$C \to 0$	$R_3 \to R_2$, $\varepsilon_3 \to \varepsilon_2$	$R_3 \to 0$, $\varepsilon_3 \to \varepsilon_2 + R_1 + R_2$
$C \to 1$	$R_3 \to R_2 + \varepsilon_2$, $\varepsilon_3 \to 0$	$R_3 \to R_1 + \varepsilon_2$, $\varepsilon_3 \to R_2$

Benchmarks: Gardiner and Ma Models

- Ma's model 1
  - Exact match for manual proofread subsystem
  - Reproduced order for redundancy subsystem without use of multiple paths
- Gardiner
  - Exact match

Prediction

$$R_n = R_{n-1} - R_{n-1}G_n + R_{n-2}G_{n-1}C_n + \varepsilon_{n-1}C_{nL}$$
$$\varepsilon_n = \varepsilon_{n-1} - \varepsilon_{n-1}C_{nL} + R_{n-2}G_{n-1}(1-C_n) + R_{n-1}G_n$$

Maximum possible Outgoing Data Accuracy is achieved when G → 0 and C → 1

$$R_{MAX} = R_{n-1} + \varepsilon_{n-1}$$
$$\varepsilon_{MIN} = 0$$

Trivial, but a valid boundary check.

Estimation in Absence of Error Correction Metrics

$$R_n = R_{n-1} - R_{n-1}G_n + R_{n-2}G_{n-1}C_n + \varepsilon_{n-1}C_{nL}$$
$$\varepsilon_n = \varepsilon_{n-1} - \varepsilon_{n-1}C_{nL} + R_{n-2}G_{n-1}(1 - C_n) + R_{n-1}G_n$$

Conservative estimate, of best Outgoing Accuracy assuming no correction, i.e. $C \rightarrow 0$

$$R_n = R_{n-1} - R_{n-1}G_n$$
$$\varepsilon_n = \varepsilon_{n-1} + R_{n-2}G_{n-1}(1) + R_{n-1}G_n$$

Gives a way to estimate lower bound on outgoing accuracy when metrics for correction rates are not known.

Example Application

Input
R_o = 1000 fields
ε_o = 100 errors
0.09 or 9% error rate

Three step process common in clinical research including 1) chart review (medical record abstraction), 2) data entry and 3) data cleaning. Input data stream comes from medical records with 1000 accurate fields and 100 fields in error.

Task	Error Generation Rate (G_i)	Error Correction Rate (C_i)	Latent Error Correction Rate (C_{iL})	Outgoing Number of Accurate Fields (R_i)	Outgoing Error Number* (ε_i)
Chart review	0.03	0.01	--	971	99
Data entry	0.0025	0	0	969	129
Cleaning	0.0025	0.01	0.01	968	132

*delayed accumulation due to modeling error generation in one step as input to error correction of next step.
Shaded areas are input to the model.

Notice that for G and C indicative of our industry, as data processing steps are added:
- Outgoing number of accurate fields decreases
- Outgoing error number increases

Limitations of the Model

- Error generation rates must be known or estimated to use the model
- Model can provide conservative estimate assuming no correction
- Model does not account for processes that selectively correct one type of error over another, all errors are treated equally with a correction rate Ci.
- Delayed accumulation of outgoing error rate due to modeling error generation in one step as input to error correction of next step.
- Model can be used for data processing where portions of data are subject to different processes. However, the complexity may become prohibitive.
- Latent error correction is accounted for by assuming one correction rate applies uniformly to all latent errors.
  - Reality of differential correction is not accounted for
  - Even the uniform correction adds an additional correction rate term which must be known or estimated.
- Model does not accept distributional input
- Dichotomous handling of data accuracy, i.e. either correct or incorrect

Strengths of the Model / Approach

- Understood from first principles
- Generalized and can be applied to common data collection and processing methods
- Takes into account error generation and correction capabilities of each data processing step
- Takes into account branching processes
- Computationally simple, calculator / spreadsheet ready
- Uses information a practitioner can easily obtain

Significance & Conclusion

1. Error rates for parts of a data handling process can be combined to estimate the outgoing error rate.
2. The method for doing such is counterintuitive and more complex than just adding or multiplying the error rates.
3. Knowledge of the published error rates, or an organization's error rates can facilitate choosing the process paths and technology that will yield appropriate quality.
4. Model provides theoretical basis and clarity for important aspects of data quality work, e.g.
   - Error prevention is more effective than error clean-up
   - Redundancy is an effective data cleaning method (Helms, Ma)
   - The more steps, the higher the error, unless steps have ability to clean errors from preceding processes

Acknowledgements

This work was supported by the Clinical and Translational Science Awards (CTSA) to Duke University (1UL1 RR 024128) and to the University of Texas Health Science at Houston (1 UL1 RR 02414).