HHS Public Access

Author manuscript

Med Care. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Med Care. 2016 October ; 54(10): e65–e72. doi:10.1097/MLR.0000000000000108.

Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction

Brock Polnaszek, BS[1,2], **Andrea Gilmore-Bykovskyi, RN, MS**[4,1], **Melissa Hovanes, RN**[1], **Rachel Roiland, PhD, RN**[2], **Patrick Ferguson, MPH, CHES**[3], **Roger Brown, PhD**[4], and **Amy JH Kind, MD, PhD**[1,2]

[1]Department of Medicine, Geriatrics Division, University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin

[2]Geriatric Research Education and Clinical Center (GRECC), William S Middleton Hospital, United States Department of Veterans Affairs, Madison, Wisconsin

[3]University of Wisconsin, Department of Population Health Sciences

[4]University of Wisconsin School of Nursing

Abstract

Background—Unstructured data encountered during retrospective electronic medical record (EMR) abstraction has routinely been identified as challenging to reliably abstract, as this data is often recorded as free text, without limitations to format or structure. There is increased interest in reliably abstracting this type of data given its prominent role in care coordination and communication, yet limited methodological guidance exists.

Objective—As standard abstraction approaches resulted in sub-standard data reliability for unstructured data elements collected as part of a multi-site, retrospective EMR study of hospital discharge communication quality, our goal was to develop, apply and examine the utility of a phase-based approach to reliably abstract unstructured data. This approach is examined using the specific example of discharge communication for warfarin management.

Research Design—We adopted a “fit-for-use” framework to guide the development and evaluation of abstraction methods using a four step, phase-based approach including (1) team

Corresponding Author: Amy JH Kind, MD, PhD, 2500 Overlook Terrace, William S Middleton VA GRECC, Madison, WI 53705, Fax: 608.280.7291, Tel: 608.280.7000, ajk@medicine.wisc.edu. Alternate Corresponding Author: Andrea Gilmore-Bykovskyi, algimore@wisc.edu.

Complete Author Information: Brock Polnaszek, BS, 800 University Bay Drive, Suite 210, Madison, WI 53705, Tel: 608.262.8310, Email: polnaszek@wisc.edu

Andrea Gilmore-Bykovskyi, RN, MS, University of Wisconsin School of Nursing, K6/117 Clinical Science Center, 600 Highland Avenue, Madison, WI 53792, Tel: 608.262.3057, Email: algilmore@wisc.edu

Melissa Hovanes, RN, 800 University Bay Drive, Suite 210, Madison, WI 53705, Tel: 608.262.8310, Email: hovanes@wisc.edu

Rachel Roiland, PhD, RN, 2500 Overlook Terrace, William S Middleton VA GRECC, Madison, WI 53705, Tel: 608.256.1901, Email: raroiland@wisc.edu

Patrick Ferguson,MPH, CHES, 800 University Bay Drive, Suite 210, Box 9445, Madison, WI 53705, Tel: 608.265.8784, Email: pdferguson@wisc.edu

Roger Brown, PhD, Clinical Science Center H6/273, 600 Highland Avenue, University of Wisconsin, Madison, WI 53792, Tel: 608-263-5281, Email: rlbrown3@wisc.edu

Amy JH Kind, MD, PhD, 2500 Overlook Terrace, William S Middleton VA GRECC, Madison, WI 53705, Fax: 608.280.7291, Tel: 608.280.7000, email: ajk@medicine.wisc.edu

building, (2) identification of challenges, (3) adaptation of abstraction methods, and (4) systematic data quality monitoring.

Measures—Unstructured data elements were the focus of this study, including elements communicating steps in warfarin management (e.g., warfarin initiation) and medical follow-up (e.g., timeframe for follow-up).

Results—After implementation of the phase-based approach, inter-rater reliability for all unstructured data elements demonstrated kappas of ≥ 0.89 -- an average increase of + 0.25 for each unstructured data element.

Conclusions—As compared to standard abstraction methodologies, this phase-based approach was more time intensive, but did markedly increase abstraction reliability for unstructured data elements within multi-site EMR documentation.

Keywords

unstructured data; medical record abstraction; communication documents

Introduction

Improving care coordination and communication between sites of care has become a major target for national health reform initiatives.[1-3] Hospital discharge communication plays a critical role in this process, by directly informing care plan development in the next setting of care.[4-6] Nevertheless, providers report that poor discharge communication is common and leads to interrupted care plans, medication discrepancies and avoidable 30-day readmissions.[4-9] Only limited standards exist to inform the creation of discharge communication, and as a result, the content and format of these communications vary considerably within and across institutions.[4-6,10] The advent and spread of electronic medical records (EMRs) has further increased this variability.[4,11-13] The lack of a standardized structure results in most discharge communications being composed of primarily free-text or “unstructured” data.

Unstructured data is information that is documented without standard content specifications, often recorded as free text.[14,15] In contrast, structured data is generally entered into discrete data fields with standardized responses or parameters (e.g. age, weight). Although unstructured data is frequently used by providers to communicate plan of care components within discharge communications, it presents a considerable challenge when quality assessors or researchers need to reliably assess for the presence of those components.[14,15] Without reliable measurement of these unstructured components, it is difficult to ascertain baseline status or changes in communication quality. The Agency for Healthcare Research and Quality identifies this issue of unstructured data as a major barrier to quality measurement.[15]

Medical record abstraction is a method frequently employed by quality assessment teams and researchers to extract information from medical records, but standard methods primarily target structured data.[16-24] Limited research is available to inform the reliable abstraction of

unstructured data, especially when that data spans multiple EMRs with inherent variations in documentation format and structure.[4-6,14,25]

The objective of this paper is to discuss the utility of a phase-based approach for reliable abstraction of unstructured data elements in multi-site, EMR-based studies, using the specific example of abstracting unstructured discharge communications on warfarin management. The development, application and examination of this approach are outlined in the context of a National Institutes of Health (NIH)-funded multi-site, retrospective EMR-based medical record abstraction study of written discharge communications. This work is presented both as a demonstration of the complexity of rigorous assessment of clinician discharge communication and as a detailed guide to make such an assessment more manageable.

Methods

The Challenge of Unstructured Data Abstraction: An Example

As part of an NIH-funded study to assess the content, quality and impact of hospital-to-nursing home discharge communication, key components informing the patient's plan of care needed to be abstracted retrospectively from 2,079 discharge communication documents, created during the years 2003-2008 and housed within separate EMRs of two study hospitals. Multiple communication components, both structured and unstructured, were abstracted from these documents to evaluate discharge communication quality. For example, expert-recommended warfarin communication unstructured data elements were examined including the presence or absence of communications as to: (1) whether warfarin was initiated or (2) changed during the present admission, and (3) documentation that the responsible party for post-hospital warfarin management was contacted. Additionally, the presence or absence of communications as to: (1) timeframe for medical follow-up and (2) instructions for how medical follow-up was to be arranged were also assessed. Although important to care, none of these components are mandated in discharge communication, so are typically included within free-text/unstructured data, if at all.

Upon study initiation in 2009, standard structured data abstraction techniques[16-24,26,27] were applied to both the structured and unstructured components of interest. Standard abstraction methodology typically consists of (1) developing abstraction protocols, tools/forms, and processes (typically without abstractor input), (2) piloting tools/forms and protocols using pre-selected, standardized patient record examples for abstractor training, (3) variable communication amongst members of the abstraction team following the training period, and (4) data quality and fidelity monitoring, usually with Cohen's Kappa statistical technique.[28-30]

Although sufficient for the structured data components, these standard abstraction approaches were inadequate for unstructured data elements, especially as the project began abstracting records from the years in which the initial study hospital transitioned from a paper medical record to an EMR (See Figure 1 for an overview of study timeline and changes in data sources, years of data and abstractors). The introduction of an EMR led to changes in discharge communication document location, archiving, format and version

control. It was common to encounter multiple versions of the same document in different stages of completion within the EMR. These issues of version control in the setting of format and archiving variability led to abstractor confusion, especially since unstructured data content typically varied between versions and each version was visible to the abstractors. Although the unstructured data elements represented relatively simple concepts (i.e., directions for warfarin management and medical follow-up), clinicians could touch on these concepts in their free-text communications using a wide variety of wordings interpreted differently by different abstractors. As a result, random re-abstractions resulted in lower than expected inter-rater reliability (IRR) statistics (<0.60) and percent agreement ratings for several unstructured data elements. As the research progressed to include the second study hospital, with a completely different EMR, it became clear that continued application of the standard abstraction methodology was unadvisable.

The EMR variability over the study period resulted in the identification of syntactic (i.e., difference in representation of the data elements) and semantic (i.e., differences in the meaning of data elements) variability relating to the format, structure, and location of discharge summary data elements. Key differences between and within each hospital's EMR included: (1) the format/structure of the pre-EMR paper-based medical record, (2) the staging, timing and extent of EMR-implementation for discharge communication processes, (3) changes in versions of EMR software, and (4) changes in the format/structure of discharge summary data elements over the course of the study.

Standard abstraction methodology did not account adequately for these contextual and data format challenges. Table 1 provides definitions and examples of the unstructured data and syntactic and semantic data quality variability encountered. Very little published literature addressed this topic.

To improve data quality, we adopted a “fit-for-use”[27] framework to guide the development and evaluation of abstraction methods to accommodate the variability we encountered in each EMR. Specifically, we focused on developing procedures that systematically identified, assessed and accommodated for the data variability associated with unstructured data elements and the syntactic/semantic EMR-specific differences.[27]

Development of the Phase-Based Abstraction Approach for Unstructured Data

To meet our needs, adaptations to standard abstraction methods were made and implemented in four phases (Figure 2), as below. Each phase was repeated as we entered into each EMR in the study.

Phase 1 – Team Building

Research Team Composition—We constructed a primary abstraction team made up of a geriatric physician, nurse scientist, and medical/nursing students, who as a composite multi-disciplinary team could attest to the care processes in both the hospital and the nursing home setting (i.e., the sites of interest in this study). Collectively, primary abstractors had both a clinical background and strong understanding of typical EMR structure. This prior experience allowed abstractors to identify salient variations in unstructured data elements.

Abstractors were trained using sample records and standardized tools/manuals, similar to standard structured abstraction approaches.

Research Team Empowerment—In contrast to the abstraction approaches used for structured data, we employed the primary abstractors as active informants in the research process. Under the direction of senior research team members, primary abstractors were empowered to initiate changes in study protocols and start/stop abstraction upon identification of data quality concerns, resulting in frequent interruptions and protocol modifications. This contrasts with previous abstraction methods wherein the abstraction tools and protocols are generally developed prior to abstractor training and without abstractor input.

Additionally, abstractors were not blinded to study hypothesis, as in standard structured abstraction approaches. This was necessary to ensure that abstractors were not simply following the abstraction protocol, which may have been initially flawed given the unexpected variability in context and format found in unstructured data. Instead, they were empowered to identify and provide insight and solutions to the data variability found with unstructured communication. However, study outcomes of interest were not ambiguous (i.e., death, rehospitalization) and regular random, blinded inter-abstractor reliability checks were maintained.

Empowerment and non-blinding of primary abstractors was essential in identifying important semantic variations. For example, after noticing meaningful differences in the verbiage used to describe warfarin initiation, primary abstractors identified the need for detailed decision algorithms to support abstraction of discharge communications regarding warfarin initiation during the hospital stay. (See Figure 3 for an example of decision-tree logic developed to abstract the presence/absence of discharge communication regarding warfarin initiation).

Research Team Communication—During this process, increased frequency of formal research team communication was necessary. Standard multi-site medical record abstraction studies note research team communications occurring daily during initial periods, but after the initial training is complete, communications decrease to monthly or quarterly intervals, if at all. Given the challenges associated with the syntactic, semantic, and unstructured data variability within this study, our team communicated daily for two weeks during protocol development, and then weekly during formal meetings thereafter.

Phase 2 – Identification of Challenges

Piloting the Abstraction Tool and Protocol—In order to understand the extent of data quality issues associated with the unstructured elements, we first piloted our abstraction tool and protocol with each EMR system. The goal of this pilot was to identify and assess the extent of unstructured data communication, syntactic variability within and across EMRs and semantic variability across EMRs. The pilot incorporated 50 discharge communication documents, employed the initial abstraction tool and was conducted separately by two abstractors.

Documentation of Challenges—The nature and frequency of variation in data quality were documented in a detailed abstraction log by each abstractor. Examples of items documented in the logs include: differing locations of discharge communication documents within the EMR, rationale for selection of a particular document when multiple versions of that document were available, and examples of unstructured wordings thought to convey the same basic concept (i.e. verbs used to denote warfarin initiation).

These challenges were discussed during weekly meetings, with formal decisions made so that future abstractions encountering similar scenarios could employ a standardized approach. Challenges like these were encountered frequently as the team began abstraction of a new EMR, but less often as standardized approaches were decided upon for a particular EMR/system. This tracking enabled us to adapt protocols and tools to be resistant to data complexities and variabilities, promoting stronger reliability. Figure 2 includes a description of the phase-based approach as well as reflective questions that can help guide others through this process.

Phase 3 – Adaptation of Abstraction Methods

Addressing Challenges—If a primary abstractor felt that existing tools/protocols inadequately addressed a specific scenario due to data variability, the situation was reviewed with another team member to see if a consensus could be reached. If neither the protocol nor the additional team member were enough to reach a consensus regarding the scenario, the challenge was reviewed impromptu with the entire team and the abstraction tools/protocols were modified accordingly. On occasion, syntactic variability necessitated input from others within the target hospital system. For example, there were multiple versions of certain discharge communication documents within one hospital's EMR. By contacting that hospital's medical records director, we were able to establish methods to identify the final document meant to be communicated to the next site of care, which was the one targeted for abstraction.

On-going Adaption of Tools/Protocols—Using the techniques described above, the research team adapted the existing abstraction tools and protocols to overcome the challenges with identifying the correct discharge communication document/document version for abstraction, and assessing for and systematically reviewing syntactic, semantic, and unstructured data variability within each EMR.

Document Identification—Since each EMR varied greatly, we developed new standardized document identification procedures for each EMR using on-going abstractor input. In the specific case of warfarin, the location and format of communications documenting whether the responsible party for post-hospital warfarin management had been contacted, changed within each hospital system and with each EMR implementation. Identification of this syntactic variation allowed for review and standardization of document choice procedures.

Decision Support Algorithms—The complexity of this abstraction prompted the creation of decision-support algorithms to guide abstractors for specific unstructured data

elements. Decision-tree logic pathways, such as the one for warfarin initiation (Figure 3), were created and integrated into the abstraction protocols. To create these pathways, the research team used their clinical judgment, known limitations of the data, and identified challenges from phase 2, to build a pathway that made clinical sense and could be uniformly applied to each patient. Decision-tree logic pathways were reviewed by all members of the team, and benefited from external review of providers experienced in the type of documentation being studied.

Validation of Abstraction Protocols and Tools—As there is currently no gold standard to inform the assessment of clinician-to-clinician discharge communication, a panel of experienced clinicians (N=5) with regular exposure to discharge communication in the EMR was convened to review the abstraction protocols and tools. The panel participated in a modified Delphi process informed by Jones and Hunter's consensus methods for health services research.[31] The process consisted of a series of discussions wherein the consensus team was given specific examples of discharge documents with which to evaluate the accuracy of the protocols in capturing discharge communication. While several questions were raised throughout the process, there were no substantive changes to the developed protocols. Consensus was strong that the developed tools appropriately documented the specific data elements within clinician discharge communication.

Phase 4 – Integration of Systematic Data Quality Monitoring

Assessing Reliability During Full Abstraction—Following phase 3, the modified abstraction tool and protocols were implemented at a faster pace and abstractors began full abstraction of the targeted sample with a systematic data quality monitoring process in place. Individual abstractors were assigned batches of 100 randomized patients. At the same time that one abstractor was assigned a batch of 100, a second abstractor was assigned a 10% random re-abstraction. Although abstractors were not blinded to one another's work in phases 1-3, which differs from standard abstraction methods, they were fully blinded in phase 4. As in standard abstraction, this blinding helps to ensure that any changes in reliability are due to the abstraction method. Delaying blinding until this point allowed for a fully developed and extensively tested abstraction methodology to be implemented.

Inter-rater Reliability (IRR) Assessments and Review—After completion of each batch of 100 cases and the accompanying 10% random re-abstraction, IRR and data quality were assessed using contingency tables, IRR statistics including Cohen's kappa, Gwet's AC_1[32], Brennan-Prediger[28], and percent agreement calculations. We added Gwet's[32] and Brennan-Prediger's[28] coefficients because some data elements were unevenly distributed, leading to high percent agreements but low Kappas. This is an expected limitation of Cohen's Kappa in this type of data situation.[28-30] Abstractors were not allowed to move onto the next batch of 100 patients until the reports were reviewed. If IRR statistics were below 0.85, the research team convened to identify why abstractors disagreed and if need be, modified the abstraction methods. IRR statistics were examined for each specific data element and did not distinguish between semantic or syntactic variation as most data elements of interest contained both types of data variability.

Results

The phase-based approached, described above, led to marked improvements in abstraction reliability for unstructured data elements. This can be seen in IRR statistics that compared before (2010-2011), during (2011-2012), and after (2012-2013) implementation of the phase-based approach. Previously used structured data abstraction methods performed sub-optimally for medical follow-up (*Cohen's Kappa* range 0.60 to 0.73) and warfarin management (*Cohen's Kappa range 0.57 to 0.94*) discharge communication elements. After the phase-based abstraction implementation, all five unstructured data elements of interest achieved kappa ≥ 0.89. Figure 4 demonstrates the increase in reliability noted during each stage of implementation of the phase-based approach, and Appendix Table 1(Supplemental Digital Content 1, http://links.lww.com/MLR/A726) denotes inter-rater reliability statistics for each data element over the course of implementation. As compared to structured abstraction methods, this phase-based approach achieved an average increase of kappa of +0.25 for each of the five unstructured data elements.

Costs and Benefits of the Phase-Based Approach

Development of abstraction protocols/tools using the phase-based approach resulted in an additional 2 weeks of up-front effort beyond that of traditional abstraction methods. The main benefit of this intensive 2 week period was that it resulted in protocols that limited the amount of interpretation individual abstractors needed to make while performing final abstractions of unstructured data. During this 2 week period, abstractors also dedicated time to investigating each EMR and establishing contact persons at each hospital with detailed knowledge of their EMR and discharge communication processes. These contacts helped the study team identify the correct discharge documents for abstraction, the location of different data elements in the EMR, how the paper-based record was integrated into the EMR, as well as implementation timing of each EMR.

Discussion

This phase-based method for multi-site EMR abstraction was developed to improve the reliability of abstraction of unstructured data elements. Using this approach we were able to systematically identify and address semantic and syntactic data variability found within discharge communication documents across EMR systems. As compared to standard abstraction methodologies, this phase-based approach allowed us to increase our abstraction reliabilities for unstructured data.

Care coordination and communication documents are of increasing interest to researchers and others interested in measuring changes in communication quality. Since unstructured data comprises the bulk of these communications, reliable techniques for unstructured data abstraction are needed. To our knowledge, this is the first publication to describe in detail how such abstraction methods could be developed and implemented, and the first to directly address the complexity of EMR context in abstraction.

Unstructured data is very different from structured data. Abstraction methodologies need to be adapted to take this into account. Throughout the abstraction methods literature,

unstructured data is frequently discussed as challenging to reliably collect when compared to structured data.[16-18,21,22,33,34] Studies have highlighted the vast choices in terminology used by providers to describe findings of a clinical examination (i.e., bump, thickened area, hard mass, cobblestone, fibrocystic) and the need to develop priorities, limitations, and hierarchies for abstracting unstructured data.[17] To our knowledge, this is the first study to demonstrate that unstructured data can be reliably collected. The phase-based methods offered here provide a suggested guide for how to proceed with unstructured data abstractions.

The collection of unstructured data is further complicated in the context of multiple or evolving EMRs. EMR use has steadily increased as a result of the 2010 Health Information Technology for Economic and Clinical Health (HITECH) Act, which provides incentive payments to hospitals and providers who adopt and “meaningfully use” EMRs.[35,36] As such, abstracting data directly from one or more EMRs will become more commonplace, and the complexity that this context adds to the abstraction process needs to be formally addressed in any research protocol. Although EMRs can create structured data fields within clinical templates, some providers still prefer creating all or portions of their clinical documentation in an unstructured format.[15,30] Therefore, even with a highly ‘templated’ EMR, unstructured data challenges remain. As they become more sophisticated, free-text computer extraction programs like Natural Language Processing may become more widely used in EMR-based abstraction, but until that time, manual abstraction is still commonplace.[37-39]

While this phase-based approach is promising, there are several limitations to consider. The suggested modifications from standard abstraction methods are time consuming, adding approximately 2 full weeks to this study. This approach may not be feasible for every study. However, every research abstraction process is a compromise between limited time/resources and the desire to attain the highest possible data quality.[27] Ensuring data integrity may be one of the most challenging tasks in planning and implementing a research study. Phases 2 and 3 utilized in our process were time intensive, yet produced the first available high reliability approach for the abstraction of unstructured data with semantic and syntactic variability within and across EMRs. Of course, there is no way to definitively prove which specific method or phase directly led to the increase in IRR over the course of the study, or if some other unrecognized factor drove these changes (albeit unlikely in our view). This method would benefit from further testing, but does show promise as an option for abstraction of highly challenging unstructured data in a variable and complex EMR context. As in all research, but especially in investigations of unstructured data, investigators and research teams need to consider the limitations of the data and data source in order to ensure that conclusions made during final reporting are correctly and appropriately interpreted to inform future interventions and research needs.

Given the increasing attention and interest on care coordination and communication, the need to reliably abstract unstructured data across EMRs and health systems will only grow. The phase-based approach is a conceptual framework to help investigators design their abstraction process to capture unstructured data variability with minimal sacrifice of data quality and integrity. When followed, this process is time intensive, but shows promise for offering excellent data reliability.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Received: National Institute on Aging Paul B. Beeson Patient-Oriented Research Career Development Award (K23AG034551 [PI Kind]), in partnership with the American Federation for Aging Research, the John A. Hartford Foundation, the Atlantic Philanthropies, and the Starr Foundation and the Madison VA Geriatrics Research, Education and Clinical Center (GRECC-Manuscript #2013-19). Additional support was provided by the University of Wisconsin School of Medicine and Public Health's Health Innovation Program; the Community-Academic Partnerships core of the University of Wisconsin Institute for Clinical and Translational Research; and the Clinical and Translational Science Award program of the National Center for Research Resources, National Institutes of Health (1UL1RR025011).

References

1. All-Payer Claims Database Council. [Accessed July 1, 2013] Fact Sheet: APCD and Health Reform. Jun. 2011 Available at: http://apcdcouncil.org/sites/apcdcouncil.org/files/APCD%20and%20Health%20Reform%20Fact%20Sheet_FINAL_0.pdf
2. National Transitions of Care Coalition. [Accessed: July 1, 2013] Improving Transitions of Care with Health Information Technology. 2010. Available at: http://www.ntocc.org/Portals/0/PDF/Resources/HITPaper.pdf
3. U.S. Department of Health and Human Services. [Accessed July 1, 2013] Accountable Care Organizations: Improving Care Coordination for People with Medicare. Available at: http://www.healthcare.gov/news/factsheets/2011/03/accountablecare03312011a.html
4. King BD, Gilmore-Bykovskyi A, Roiland R, et al. The Consequences of Poor Communication During Hospital-to-Skilled Nursing Facility Transitions: A Qualitative Study. J Am Geriatr Soc. 2013
5. van Walraven C, Mamdani M, Fang J, et al. Continuity of care and patient outcomes after hospital discharge. J Gen Intern Med. 2004; 19(6):624–631. [PubMed: 15209600]
6. Kind, A.; Smith, M. AHRQ Patient Safety: New Directions and Alternative Approaches vol 2: Culture and Redesign. Rockville, MD: Agency for Healthcare Research and Quality; 2008. Documentation of Mandated Discharge Summary Components in Transitions from Acute to Sub-Acute Care; p. 179-188.
7. van Walraven C, Laupacis A, Seth R, et al. Dictated versus database-generated discharge summaries: a randomized clinical trial. CMAJ. 1999; 160(3):319–326. [PubMed: 10065073]
8. Kind A, Anderson P, Hind J, et al. Omission of dysphagia therapies in hospital discharge communications. Dysphagia. 2011; 26(1):49–61. [PubMed: 20098999]
9. Kind AJ, Thorpe CT, Sattin JA, et al. Provider characteristics, clinical-work processes and their relationship to discharge summary quality for sub-acute care patients. J Gen Intern Med. 2012; 27(1):78–84. [PubMed: 21901489]
10. Joint Commission on the Accreditation of Healthcare Organizations (JCAHO). [Accessed July 1, 2013] Standard IM.6.10, EP 7. Available at: http://www.jointcommission.org/NR/rdonlyres/143FDA42-A28F-426D-ABD2-EAB611C1FD24/0/C_HistoryTracking_BHC_RC_20090323v2.pdf
11. Desroches CM, Audet AM, Painter M, et al. Meeting meaningful use criteria and managing patient populations: a national survey of practicing physicians. Ann Intern Med. 2013; 158(11):791–799. [PubMed: 23732712]
12. Walker JM. Electronic medical records and health care transformation. Health Aff (Millwood). 2005; 24(5):1118–1120. [PubMed: 16162552]
13. Howard J, Clark EC, Friedman A, et al. Electronic health record impact on work burden in small, unaffiliated, community-based primary care practices. J Gen Intern Med. 2013; 28(1):107–113. [PubMed: 22926633]

14. Mishuris RG, Linder JA. Electronic health records and the increasing complexity of medical practice: “it never gets easier, you just go faster”. J Gen Intern Med. 2013; 28(4):490–492. [PubMed: 23247583]
15. Agency for Healthcare Research and Quality. [Accessed July 1, 2013] Prospects for care coordination measurement using electronic data sources: Challenges of measuring care coordination using electronic data and recommendations to address those challenges. Available at: http://www.ahrq.gov/qual/prospectscare/prospects1.htm
16. Reisch LM, Fosse JS, Beverly K, et al. Training, quality assurance, and assessment of medical record abstraction in a multisite study. Am J Epidemiol. 2003; 157(6):546–551. [PubMed: 12631545]
17. Eder C, Fullerton J, Benroth R, et al. Pragmatic strategies that enhance the reliability of data abstracted from medical records. Appl Nurs Res. 2005; 18(1):50–54. [PubMed: 15812736]
18. Pan L, Fergusson D, Schweitzer I, et al. Ensuring high accuracy of data abstracted from patient charts: the use of a standardized medical record as a training tool. J Clin Epidemiol. 2005; 58(9): 918–923. [PubMed: 16085195]
19. Liddy C, Wiens M, Hogg W. Methods to achieve high interrater reliability in data collection from primary care medical records. Ann Fam Med. 2011; 9(1):57–62. [PubMed: 21242562]
20. Simmons B, Bennett F, Nelson A, et al. Data abstraction: designing the tools, recruiting and training the data abstractors. SCI Nurs. 2002; 19(1):22–24. [PubMed: 12510501]
21. Kung HC, Hanzlick R, Spitler JF. Abstracting data from medical examiner/coroner reports: concordance among abstractors and implications for data reporting. J Forensic Sci. 2001; 46(5): 1126–1131. [PubMed: 11569554]
22. Engel L, Henderson C, Fergenbaum J, et al. Medical record review conduction model for improving interrater reliability of abstracting medical-related information. Eval Health Prof. 2009; 32(3):281–298. [PubMed: 19679636]
23. Engel L, Henderson C, Colantonio A. Eleven steps to improving data collection: Guidelines for a retrospective medical record review. Occupational Therapy Now. 2008; 10:17–20.
24. Allison JJ, Wall TC, Spettell CM, et al. The art and science of chart review. Jt Comm J Qual Improv. 2000; 26(3):115–136. [PubMed: 10709146]
25. van Walraven C, Duke SM, Weinberg AL, et al. Standardized or narrative discharge summaries. Which do family physicians prefer? Can Fam Physician. 1998; 44:62–69. [PubMed: 9481464]
26. Banks NJ. Designing medical record abstraction forms. Int J Qual Health Care. 1998; 10(2):163–167. [PubMed: 9690890]
27. Kahn MG, Raebel MA, Glanz JM, et al. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012; 50(Suppl):S21–29. [PubMed: 22692254]
28. Brennan RL, Prediger DJ. Coefficient Kappa - Some Uses, Misuses, and Alternatives. Educational and Psychological Measurement. 1981; 41(3):687–699.
29. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990; 43(6):551–558. [PubMed: 2189948]
30. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43(6):543–549. [PubMed: 2348207]
31. Jones J, Hunter D. Consensus methods or medical and health services research. BMJ. 1995; 311(7001):376–380. [PubMed: 7640549]
32. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008; 61(Pt 1):29–48. [PubMed: 18482474]
33. Hayward RA, Hofer TP. Estimating hospital deaths due to medical errors: preventability is in the eye of the reviewer. JAMA. 2001; 286(4):415–420. [PubMed: 11466119]
34. Stein HD, Nadkarni P, Erdos J, et al. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. J Am Med Inform Assoc. 2000; 7(1):42–54. [PubMed: 10641962]
35. Patel V, Jamoom E, Hsiao CJ, et al. Variation in electronic health record adoption and readiness for meaningful use: 2008-2011. J Gen Intern Med. 2013; 28(7):957–964. [PubMed: 23371416]

36. [Accessed July 1, 2013] Office of the National Coordinator for Health Information Technology. Policymaking, Regulation, & Strategy: Meaningful Use. Available at: http://www.healthit.gov/policy-researchers-implementers/meaningful-use

37. Voorham J, Denig P. Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners. J Am Med Inform Assoc. 2007; 14(3):349–354. [PubMed: 17329733]

38. Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med. 1999; 74(8):890–895. [PubMed: 10495728]

39. Friedman C, Shagina L, Lussier Y, et al. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004; 11(5):392–402. [PubMed: 15187068]

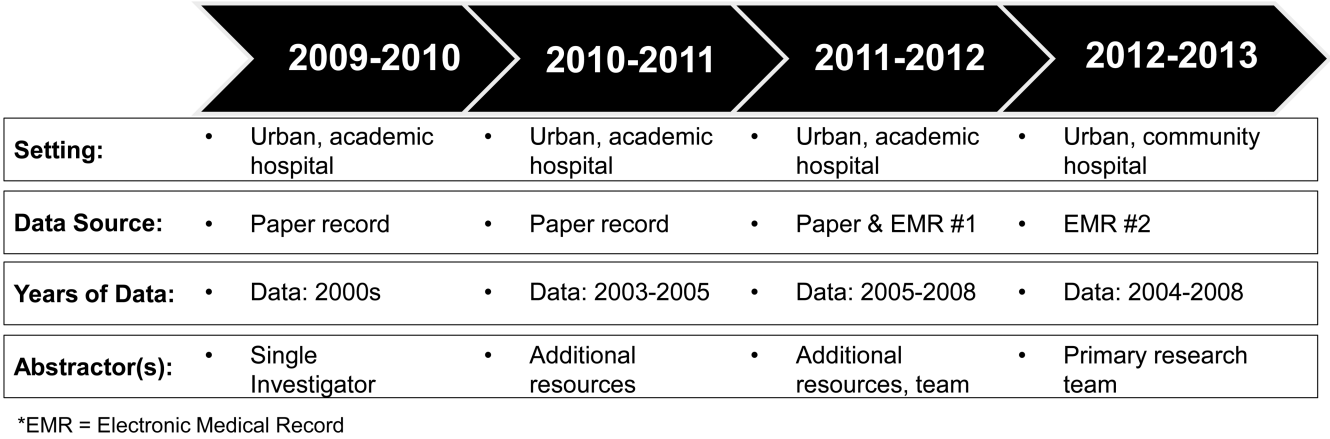

Figure 1. Study Timeline and Changes in Hospital Setting, Data Sources, Years of Data, and Abstractors

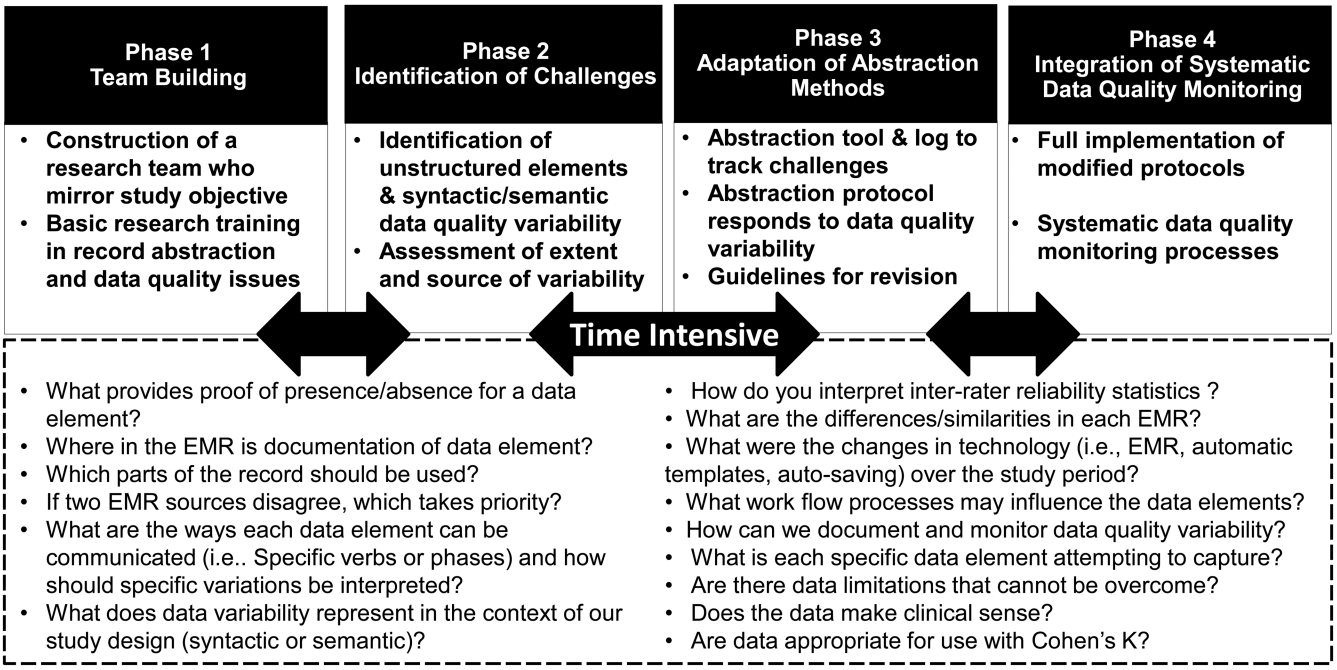

Figure 2. Implementation of Phase-Based Abstraction Approach for Unstructured EMR-Based Data

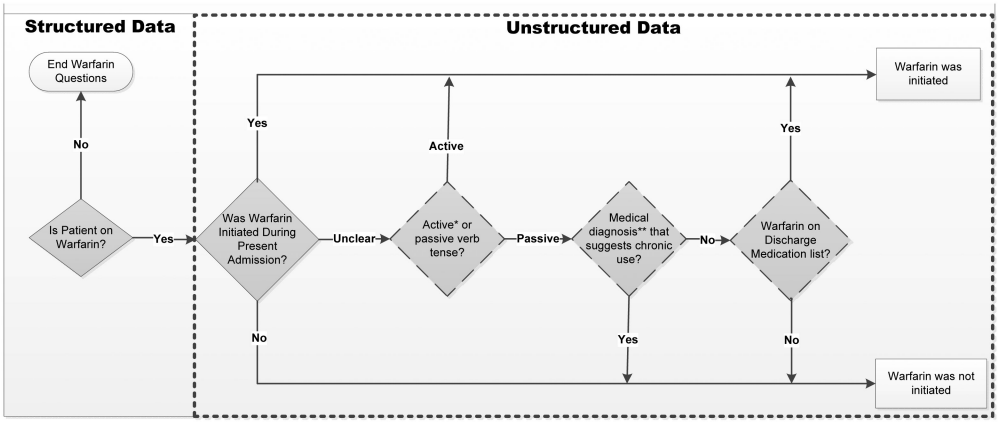

Figure 3. Example of Decision-Tree Logic for Abstraction of Structured and Unstructured Discharge Communication of Warfarin Initiation

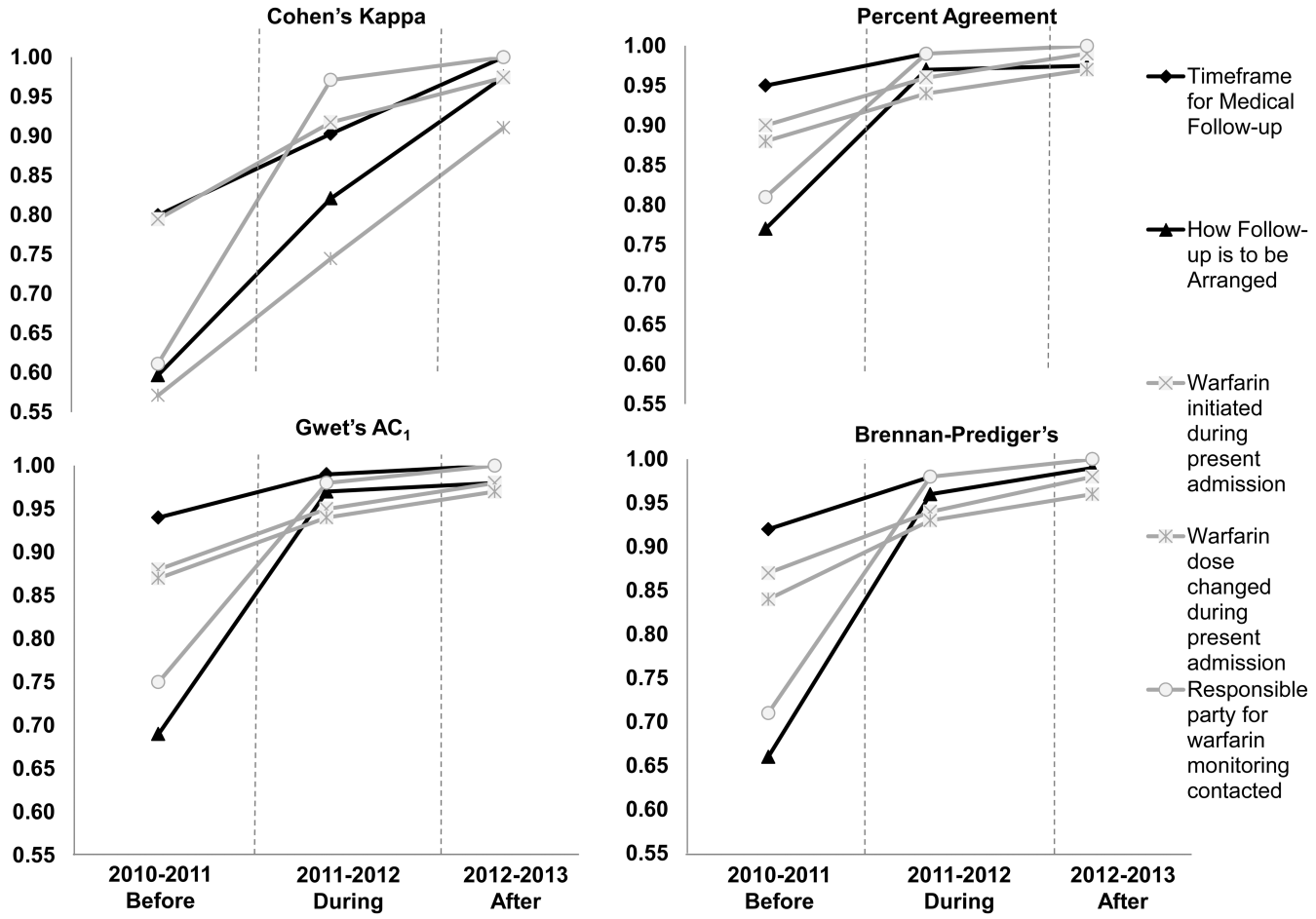

Figure 4. Change in Inter-rater Reliability Statistics through Implementation of Phase-Based Abstraction for Discharge Communications of Warfarin Management and Follow-up Instructions

Table 1

Definitions and Examples of Unstructured Data, Syntactic Variability and Semantic Variability in Discharge Communication

Category	Examples	Specific Data Elements
Unstructured Data: Generally recorded as free text, with minimal guidelines for content, format or structure of information.	Progress notes Nursing notes Discharge summaries Physical therapy, occupational therapy, and speech language pathology notes History and physical examination Admission assessments Hospital History	Verbs to Describe Warfarin Initiation U*tilized, maintained, initiated, started, deployed, observed, begun, etc.* INR Assessment Timeframe W*eeks, as needed, specific days, per next provider, to be set at skilled nursing facility, absent* *INR Goal* *Within therapeutic range, single digits like 2.0, 2.0-3.0, absent* How medical follow-up is to be arranged *Routine needed, scheduling in process, to be scheduled / determined, defer to another service / provider, communicated with patient, 2 weeks, patient will / is to / should follow-up* Medical Follow-Up with Whom and Timeframe *PCP but no PCP identified, provider at next facility, notify provider when patient arrives, if PCP feels necessary,rehab unit to arrange, 2 weeks, as needed*
Syntactic Variability: Data variability caused by differences in the representation of data elements. These issues are detectable and resolvable using single-site data	Weight, age, sex, birth date may be recorded and stored in different locations, formats or units within an EMR	Multiple written, electronic, and combination of versions with each changing format and location within EMR Service / Provider dependent format for discharge summary Warfarin dose change discussed in narrative text but absent all together on admission/discharge medication list Medical follow-up standard format (date, time, location, and with whom) PCP information automatically generated at the top of discharge summary
Semantic Variability: Data variability caused by differences in the meaning of data elements. Difficult to detect using single-site data alone because data semantics tend to be consistent within an institution	Fasting and random blood glucose, finger-stick or venipuncture, serum or plasma measurements would result in glucose values that do not represent the same concept	Physician signatures *Authenticated, hand-written signature, electronically signed, unsigned* Multiple locations for discharge summaries *Historical encounter, electronic discharge summary link on face sheet, discharge summary tab, notes, scanned PDF files* Pro-time versus INR diagnostic test Anticoagulation clinic on site and integrated in EMR process Primary Care Physician Information routinely *Located at the top of the discharge summary or manually entered by physician in signature block*

PCP = Primary Care Provider; INR = International Normalized Ratio; EMR = Electronic Medical Record