Training as an Intervention to Decrease Medical Record Abstraction Errors Multicenter Studies

Meredith Nahm ZOZUS, PhD;[a] Leslie W YOUNG, MD;[b] Alan E SIMON, MD; [c]
Maryam GARZA, MMCi; [a] Lora LAWRENCE, RN;[a] Songthip T OUNPRASEUTH,
PhD;[a] Megan BLEDSOE, MSc; [d] Sarah NEWMAN-NORLUND, MS;[e] J. Dean
JARVIS, RN, MBA; [f] Mary MCNALLY, RRT; [f] Kimberly R. HARRIS;[a] Russell
MCCULLOH, MD;[g] Rachel AIKMAN, RN, MS; [g] Sara COX, RN;[h] Lacy MALLOCH;[i]
Anita WALDEN, MS; [a] Jessica SNOWDEN, MD;[a] Irene Mangan CHEDJIEU, BDS,
MPH;[a] Chester A WICKER; [a] Lauren ATKINS;[j]Lori A DEVLIN, DO MHA[c]

[a] *University of Arkansas for Medical Sciences, Little Rock, AR;* [b] *Larner College of
Medicine at the University of Vermont, Burlington, VT;* [c]*National Institutes of Health,
Bethesda, MD;* [d]*Children's Mercy Hospital, Kansas City, MO;* [e]*Palmetto
Health,Columbia, SC;* [f]*Children's Hospital at Dartmouth-Hitchcock, Lebanon, NH*;
[g]*University of Nebraska Medical Center, Omaha, NE*; [h]*University of Montana School of
Public and Community Health Sciences, Missoula, MT*; [i]*University of Mississippi
Medical Center, Jackson, MS;* [j]*KS; University of Louisville School of Medicine,
Louisville, KY;*

Abstract: Studies often rely on medical record abstraction as a major source of data. However, data quality from medical record abstraction has long been questioned. Electronic Health Records (EHRs) potentially add variability to the abstraction process due to the complexity of navigating and locating study data within these systems. We report training for and initial quality assessment of medical record abstraction for a clinical study conducted by the IDeA States Pediatric Clinical Trials Network (ISPCTN) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) Neonatal Research Network (NRN) using medical record abstraction as the primary data source. As part of overall quality assurance, study-specific training for medical record abstractors was developed and deployed during study start-up. The training consisted of a didactic session with an example case abstraction and an independent abstraction of two standardized cases. Sixty-nine site abstractors from thirty sites were trained. The training was designed to achieve an error rate for each abstractor of no greater than 4.93% with a mean of 2.53%, at study initiation. Twenty-three percent of the trainees exceeded the acceptance limit on one or both of the training test cases, supporting the need for such training. We describe lessons learned in the design and operationalization of the study-specific, medical record abstraction training program.

Keywords. Data collection, data quality, chart review, medical record abstraction, clinical data management, clinical research informatics, clinical research

1. Introduction

We define medical record abstraction (MRA) as a process in which a human manually searches through an electronic or paper medical record to identify data required for

secondary use [1]. Abstraction involves direct matching of information found in the record to the data elements required for a study. Abstraction also commonly includes subjective tasks such as categorizing, selection of one value from multiple options, coding, interpreting, and summarizing data from the medical record as well as objective tasks such as transformation, formatting, and calculations based on abstracted data. The extent to which MRA relies on subjective tasks is variable and based on the design and operationalization of individual studies. Some studies constrain abstraction toward objective tasks while others rely to a greater extent on human interpretation and decision-making. Studies relying on data found in narrative parts of the record or on data equally likely to be charted in multiple places frequently require the latter. The subjective aspects of MRA differentiate it from other data collection methods in clinical research that are largely reliant on objective processes. Further, others have characterized medical record abstraction as having a higher cognitive load than other data collection and processing methods[2]. A pooled analysis of data quality among clinical studies relying on MRA found MRA to be associated with the highest rates of error in data collection and processing (mean 960, median 647, standard deviation of 1,018 errors per ten thousand fields)[3]. For these reasons, we gave special consideration to quality assurance and control of the MRA process.

As evidenced by historical and current regulatory guidance on the topic,[4, 5] the medical record remains a major source of data for clinical studies. Although many studies today can be conducted with electronically extracted data, smaller studies often do not have the resources to write and validate computer programs to extract data. Additionally, many data elements are not consistently collected or available in structured form and variations in definition, collection and charting of EHR data complicate manual abstraction and electronic extraction in multisite studies. Thus, Electronic Health Record (EHR) adoption does not obviate the need for MRA. Abstractors using EHRs as the data source must still manually search through the record to identify needed data values. EHR abstractors remain hampered by many of the same issues affecting abstraction from paper charts. Further, MRA remains a primary method for validating algorithms for electronically extracting data from healthcare information systems[6]. Thus, for the foreseeable future, many studies will continue to rely on data abstracted from medical records.

Given the continued reliance of clinical studies on MRA and high MRA error rates, data accuracy from MRA remains a concern. To address the concern, a recent review produced a systems theory-based framework for data quality assurance and control of MRA processes[1]. We have applied the framework on a clinical study conducted by the National Institutes of Health, Environmental influences on Child Health Outcomes (ECHO) Program's IDeA States Pediatric Clinical Trial Network (ISPCTN)[7] and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Neonatal Research Network (NRN) and report results from the initial phase of that implementation.

2. Background

2.1. Data Accuracy from the Medical Record Abstraction Process

Data have been abstracted from medical records since the earliest days of medical record keeping. Concern about quality of data abstracted for secondary use from medical records dates back at least to the year 1746 [8].

Many have questioned the accuracy of data abstracted from medical records. As early as 1969, researchers described the association of MRA with poorly explained processes, inconsistency, and error[9, 10]. Notable reports have continued to question medical records as a data source [11-14] and recent reviews have confirmed persistence of error rates that call into question the fitness of data abstracted from medical records for secondary use [3, 15, 16].

To detect and correct abstraction errors in clinical trials regulated by the United States Food and Drug Administration, clinical trial monitors routinely visit study sites and compare data on the collection form to the medical record in a process called source data verification. However, MRA data error rates are usually not measured in this process. Thus, they are neither documented nor presented with the analysis. In general, risk-based monitoring [17] decisions about which sites, cases, and data to monitor are made *sans* quantitative data quality evidence. Likewise, MRA error rates are rarely reported in secondary analysis studies. In a recent review, less than 9% of included studies reported an error or discrepancy rate, 0% of the included studies stated the source of data within the medical record, half failed to mention abstraction methods or tools, and only 42% mentioned training or qualification of the abstractors [1]. Thus, Feinstein et al.'s 1969 intimation that medical record reviews are, "governed by the laws of laissez faire" remains characteristic [10].

2.2. The Medical Record Abstraction (MRA) Quality Assurance Framework

Zozus, et al. first introduced the MRA Quality Assurance and Control framework in 2015 and provided evidence-based guidelines for ensuring the accuracy of data abstracted from medical records [1]. The framework treats MRA as a system with inputs and outputs. Some inputs to the MRA system such as abstractor training and abstraction guidelines are controllable while others such as the quality and content of the medical record are not. Research teams use ongoing measurement of MRA system outputs, i.e., data accuracy or some surrogate thereof, to indicate when changes to controllable inputs are needed to achieve a more desirable result. A system is considered capable when changes in controllable inputs are sufficient to achieve the desired result – some systems by design are not capable. In the case of a clinical study reliant on abstraction, lack of capability occurs when the data in the medical record are frequently missing, highly variable, or so poorly documented that they cannot be reliably or accurately abstracted.

The systems theory-based framework describes two essential mechanisms to achieve the desired data accuracy from MRA: (1) quality assurance — prospective actions taken such as abstractor training, standard procedures, and job aids to assure adequate accuracy, and (2) quality control — measurement of error or discrepancy rates and use of the measurements to guide adjustments to controllable inputs to the abstraction process such as abstraction tools, procedures, and training. The framework describes four areas where *a priori* activities to assure data accuracy should be

considered: 1) choice of the data source within the medical record, 2) abstraction methods and tools, 3) abstraction environment, and 4) abstractor qualification and training.

In the MRA system, re-abstraction by the same or different person to identify discrepancies, intra- or inter-rater reliability respectively, is used as a surrogate for data accuracy. The feedback, then, consists of data discrepancies that are reported to the abstractors or otherwise used to inform changes to the controllable inputs to decrease the discrepancy rate. Thus, the re-abstraction, actually a measure of reliability, is used as a quantitative and surrogate indicator of data accuracy as feedback to control the abstraction error rate. Use of intra- or inter-rater reliability as a surrogate measure of accuracy is not without consequence. In an area such as MRA, where the underlying data source is characterized by inherent variability and uncertainty, it is possible to force consistency at the expense of accuracy.

In this study, we aim to demonstrate, describe and report operational challenges encountered in implementing initial abstractor training and assessment as a quality assurance measure in clinical study relying on abstracted data.

3. Methods

This observational and empirical study was conducted in the context of the Advancing Clinical Trials in Neonatal Opioid Withdrawal Syndrome Current Experience: Infant Exposure and Treatment study (ACT NOW). Briefly, the ACT NOW CE study is an observational study of clinical practice in Neonatal Opioid Withdrawal Syndrome (NOWS). The ACT NOW CE study is conducted in units across all levels of neonatal care from well born nurseries to intensive care units at thirty clinical sites across the United States. ACT NOW CE data are collected through medical record abstraction performed locally at participating sites. The study is being conducted to characterize and quantify variability in current clinical practice, and to identify associations to be tested in future studies — all toward improving NOWS treatment outcomes. To assure data quality, training in the medical record abstraction process for the study was designed and implemented. Timing of the training and short study start-up time drove differences in planned versus delivered training.

Training as Designed: The planned training included the following. (1) A pre-training exercise in which sites used the ACT NOW CE abstraction guidelines to abstract two cases from their EHR. (2) A didactic portion where the trainer facilitated discussion of challenges sites faced in the abstraction of the two pre-training cases and then walked through a complete abstraction of a training case. The training was assessed through (3) an assignment to independently abstract two standardized training cases. Trainee abstraction of the two standardized training cases was reviewed by the data coordinating center. The number of errors was to be counted and feedback on errors was to be provided to the trainee. If the acceptance criterion was exceeded, the trainee was to be provided with two additional standardized cases to abstract. If the acceptance criterion was not be met within six cases, the plan was to seek a different person at the site to serve as an abstractor.

Training as Delivered: The training was altered in response to operational challenges. As delivered, the training contained no pre-training exercise (1 above) and included the following. (2) The didactic portion of the training consisted of the trainer reviewing the abstraction form with trainees and discussing anticipated abstraction

challenges. This included demonstration of abstraction for key data using a standardized training case. For ISPCTN abstractors, because the ISPCTN is a new network, the training was assessed through (3) an assignment to independently abstract two standardized cases as stated above. Similarly, the abstraction for the two standardized cases was reviewed and the number of errors counted. Trainees were told the number of errors but did not receive individual feedback regarding the errors they made. If the acceptance criterion was exceeded, the trainee was provided one, rather than the planned two, additional standardized cases to abstract.

To assure transfer of training, the training was followed by an independent re-abstraction of the first three ACT NOW CE study cases. Discrepancies between the original abstraction and the independent re-abstraction were programmatically identified through the web-based Electronic Data Capture (EDC) system for the study. The same acceptance criterion used in the training was applied (an average of less than four discrepancies per case, fewer than twelve discrepancies across the three re-abstracted cases). Following the re-abstraction, a call was held to review the discrepancies. The independent local abstractors were also required to attend the medical record abstraction training and meet the acceptance criterion.

Because large differences in error rate calculations have previously been demonstrated in the literature,[18, 19] we use the error/discrepancy rate calculation method described in the Good Clinical Data Management Practices (GCDMP)[20] and similarly defined an abstraction error or discrepancy as any meaningful difference in the abstracted value and the standard or re-abstracted value unless ambiguity in the abstraction guidelines was found to be the cause. Further, we calculated error/discrepancy rates using all fields on the abstraction form as well as using only populated fields on the form. The latter uses only fields populated with data as the denominator and provides a conservative point estimate of the error/discrepancy rate. Calculating both "all field" and "populated field" rates provides both an optimistic and a conservative measurement.[19] Data from training cases were documented to evidence study training and abstractor qualification for compliance with Good Clinical Practice.[4]

This study of a training intervention (IRB#217927) consisted of (1) secondary use of EHR data at our institution to create the seven standardized and redacted training cases based on real patients with NOWS cared for in a neonatal intensive care unit, and (2) secondary analysis of abstraction error rates on the cases used in training. This training study received a determination of not human subject research as defined in 45 CFR 46.102 by the University of Arkansas for Medical Sciences Institutional Review Board (IRB).

A confidence interval approach was used to assess point estimates of the error rate. The ACT NOW CE study data collection form contains 258 fields. To set the acceptance criterion, we conservatively estimated 316 relevant fields for two abstracted cases. If eight errors are found across the two standardized training cases (2.53% error rate), the corresponding 95% confidence interval using Pearson's exact is (1.10, 4.93). The goal of training was to achieve an error rate for each abstractor of no greater than 4.93%. Error rates were assessed at both the abstractor level and at the overall, across all abstractors. Error rates are normalized to errors per 10,000 fields for comparisons across cases having differing numbers of fields. The actual field count for the ACT NOW CE abstraction form is 317 fields. This field count was used for the "all field" error rate. The denominator for the error rate calculation is the number of trainees (69) multiplied by the number of fields on the abstraction form for a total denominator of

21,873 fields. The actual number of fields populated on the standardized cases was, 64 and 71 fields for cases one and two respectively. The denominators for the "populated field error rates" were similarly calculated using these field counts.

4. Results

4.1 Error rate results

The error rate across all abstractors (Table 1) fell within expectations and was below the average MRA discrepancy rate reported in a recent pooled analysis[3]. In total, sixty-nine site abstractors including those designated as re-abstractors for the quality control process were trained. Two hundred and eight errors occurred on training case 1 across all abstractors and three hundred and fifty-five errors occurred on training case 2 across all abstractors. The acceptance criterion for individual abstractors was exceeded twice for case 1 and sixteen times for case 2. Of note, the patient in case 1 did not receive any pharmacologic treatment whereas case 2 did receive pharmacologic treatment. The latter pharmacologic case was designed to be more difficult to abstract. As expected, abstractors considered it so.

Table 1: Overall Error Rate Estimate Range

Training Case	Errors	All Field Error Rate[*] (95% Conf. Interval)	Populated Field Error Rate[*] (95% Conf. Interval)
1	208	95 (83, 109)	471 (410, 538)
2	355	162 (146, 180)	725 (654, 801)

*Error rates are normalized to errors per 10,000 fields.

Overall, 16 (23%) of the trainees exceeded the acceptance limit - greater than eight errors across both of the standardized training cases (Figure 1). These trainees went on to complete one or more additional standardized training cases. Figure 1 shows the distribution of number of trainees with each error count in case 1, case 2, and across both training cases combined. Both the mean and range of the number of errors per case were higher in case 2.

Solicited and informal responses from trainees, data coordinating center staff, and study leadership in face-to-face meetings, email, and teleconference settings were generally supportive of the assertion that the abstraction training was helpful. The coordinating center found the training helpful in setting the expectation for consistency and revealing areas of ambiguity in the abstraction guidelines as well as differences in clinical process and documentation across the sites. However, development and provision of the training also encountered challenges. These challenges and feedback from the aforementioned stakeholders informed updates to the abstraction guidelines and data collection form.

4.2 Ambiguity and gaps in the abstraction guidelines identified through the training

The training identified instances of inconsistency and ambiguity in the abstraction guidelines. The training surfaced several examples of data commonly documented in multiple places in the EHR such as values charted in flow sheets as well as progress
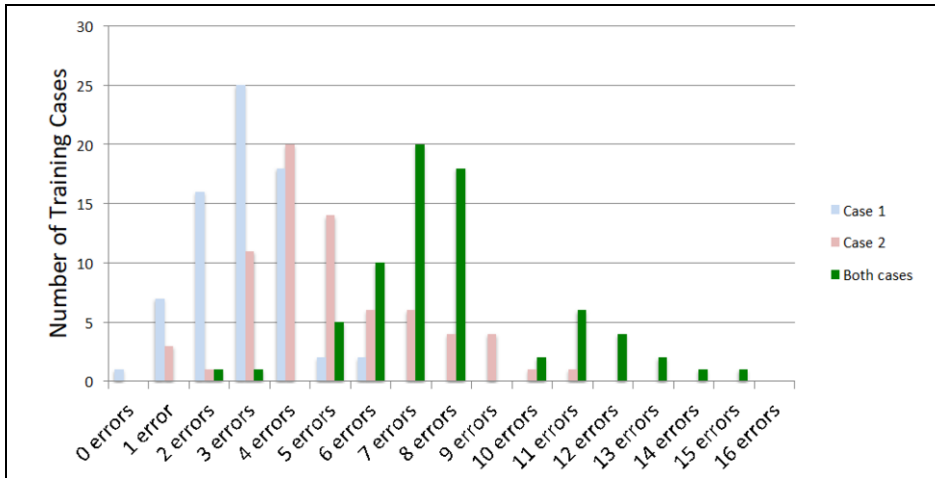
Figure 1: Frequency Distribution: Errors Across Trainees for the Two Training Cases
The "Both cases" category in the figure is the sum of errors across both cases by trainee.

notes and summaries. Usually, there was a preferred source or location for the information. For example, non-pharmacologic support information from the flow sheet was deemed closer to the original capture and more contemporaneous with the time of observation and therefore preferred over the same information that may also be documented in nursing or physician notes. Charting practices varied by site. Another site-to-site difference that the training surfaced concerned differences in receipt and documentation of prior care for babies transferred from other facilities. Babies transferred into some study site facilities are accompanied by paper records and typically lacked maternal information. When sent, these paper records are scanned and uploaded into the EHR by some facilities. Scanned records are harder to find and information contained within are usually not directly searchable, necessitating reading every page in order to abstract the case. Further, some participating facilities transferred the babies back to the referring institution upon discharge. Transfer in and transfer out of the site's facility led to, at a minimum, more tedious and time consuming abstraction, and in some cases, inaccuracies in the data abstraction process and missing data. The research team identified that it was unclear to abstractors whether transfers to other institutions should be considered discharges or not despite this being defined in the abstraction guidelines. The training also surfaced differences in the definition of nonpharmacologic care and discharge planning processes. Finally, the training surfaced variability in EHR implementation and use at sites with respect to the percentage of the study data found in structured form versus text notes. Additional examples are enumerated in the appendix.

4.3 Operational challenges encountered in delivering abstraction training

We also encountered challenges delivering the training as designed. Differences between the planned training and the training as delivered are described in the methods section. The example abstraction of key data in the didactic portion of the training and the standardized training cases were based on the Epic EHR and were not as helpful to

sites using other EHRs. Further, the standardized training cases used real and redacted data; abstracting from redacted cases was not reflective of abstraction in the real world. The team manually creating and reviewing the standardized training cases did so by pulling data from multiple patients; trainees detected errors in the initial training cases such as more than one gender being listed and babies' weights remaining static for multiple days. These errors distracted and confused trainees and emphasize the importance of careful test case preparation for MRA training. Following the training, due to time constraints, the coordinating center team provided written feedback involving the trainees' score and common errors, rather than detailed individual feedback for the trainee or site to reinforce the abstraction guidelines. An additional implementation challenge concerned the timing of the training. The start-up time was very compressed and successful completion of the training was required prior to starting the study at IDeA State sites. The training was designed with a pre-training exercise for sites to abstract two cases from their local EHR. The purpose of this activity was to ensure that sites attended the training session prepared with problems encountered using the abstraction guidelines to abstract from their local EHR; in particular, to help identify instances of the abstraction guidelines being too general or the presence of study data requirements which could not be met through the local EHR. Multiple sites were not able to do this as the training was done during site start-up. Many sites did not yet have IRB approval and were not able to use existing data for training in this way.

Multiple site study coordinators commented that the QC process following the training was more helpful than the standardized cases. The QC process consisted of local re-abstraction of real study cases using the local sites' EHR followed by individual feedback.

4.4 Resources required to develop and operationalize the training

Resources required to undertake the training were significant. Creation of the seven training cases necessitated a secondary data use application to the central IRB for the study taking five hours to write and submit. Creation of the standardized training cases took 54.25 abstraction hours plus 7 hours for double independent manual review of the redaction on the cases. Clinical operations staff at the coordinating center spent additional time preparing the training. Each of the sixty-nine trainees attended or viewed the training and spent an estimated two hours per abstraction case. Clinical operations staff at the coordinating center then spent an average of twenty-five minutes reviewing the abstracted training cases and identifying errors. Successful completion of the training was required prior to a site starting the study. To support independent abstraction for during-study quality control, two trained abstractors were required at each site. Training in abstraction added to the burden of other required trainings and occurred during the hectic start-up time period of the clinical study.

5. Discussion

Abstraction error rates of 95 and 162 errors per 10,000 fields on the two standardized training cases are favorable compared to an earlier pooled analysis (mean 960, median 647, standard deviation of 1,018 errors per ten thousand fields)[3]. We use the "all

fields" error rate because this is most comparable to how the rates for the earlier pooled analysis were calculated. Possible reasons for the favorable comparison include (1) the presence of detailed abstraction guidelines, a corresponding structured abstraction form and detailed training, (2) trainee awareness that the abstraction error rate on the training cases would be measured, and (3) the ACT NOW CE study abstraction may have been easier on average than studies in the pooled analysis. Further, the error rate on the standardized training cases is a snapshot of the first two cases for each abstractor. Thus we expect the true process capability (error rate) lies below this.

Choosing the acceptance criteria required balancing: (1) requiring an error rate that would not adversely impact study conclusions, (2) keeping the training time as low as possible, and (3) including enough cases to decrease uncertainty in the point estimate. Our choice of using two standardized training cases in the initial batch, and an acceptance criterion of eight or fewer errors across the two standardized training cases provided reasonable assurance that the error rate for each abstractor would not likely be greater than 4.93%, the upper limit of the 95% confidence interval. This acceptance criterion has face validity in that it seems obtainable given the recent pooled analysis[3] and is low enough such that we would not expect an adverse impact on the intended analysis. Further, abstraction of two training cases was estimated at study onset to take between two and four hours per case. Thus, the estimated training burden and time delay was acceptable to the study team. It is important, however, to be clear that the point estimate and associated confidence intervals achieved by this approach are for the overall abstraction error rate, i.e., over all fields on the data collection form. Achieving reasonable confidence intervals for only the subset of fields used for any specific analysis of the ACT NOW CE data would have required significantly more training cases. Indeed, error rate acceptability was based on the overall error rate and conclusions as to the acceptability of the error rate of any one data element would have considerably higher uncertainty.

Sites expressed that they would have preferred abstraction training using their in-house EHR. However three problems arise with this approach: (1) the need for short study start-up time necessitated training prior to IRB approval at a site, yet data could not be accessed at sites prior to IRB approval, (2) training on the local EHR would require either the trainer to be expert in all site EHRs or use of a local independent assessor to detect abstraction discrepancies, and (3) using local independent assessors sacrifices the ability to measure accuracy. Using standardized cases permits measurement of accuracy as opposed to reliability between two independent raters equally likely to be in error. Whereas training using local site EHRs is possible in a single site study or a study in multiple facilities with similar implementations of the same EHR, it is less feasible in a large multi-center study. Given this limitation, the need for abstraction training, and the desire for an accuracy as well as reliability assessment, we suggest the following training process.

(1) Standardized cases are prepared using an example EHR and the abstraction form.
(2) A trainer discusses and demonstrates each component of the abstraction guidelines on an example standardized training case., for site abstractors and re-abstractors.
(3) A small number of standardized training cases are distributed to trainees, abstracted by trainees and graded with feedback on errors provided to each trainee. The error rate is used to qualify abstractors and abstractors must meet a

study-specific acceptance criterion on the standardized cases prior to abstracting for the study.

(4) Following IRB approval at each site, a representative sample of charts is abstracted by the designated site abstractor and re-abstracted by an independent abstractor at the site.

(5) The abstraction and re-abstraction are compared and reviewed with the abstractor, the re-abstractor and the coordinating center. The discrepancy rate is quantified, and discrepancies and their root causes are discussed to reinforce training and improve the abstraction guidelines.

This proposed abstraction training process preserves the ability to assess accuracy and abstraction guidelines as early as possible as well as shifts as much training as possible prior to IRB approval. The process also provides for a monitored transition to abstraction from the local site EHR and ability to asses performance of the abstraction guidelines across site-to-site and EHR-to-EHR variability prior to study start.

6. Limitations

Our results are observational. We were able to standardize the abstraction process, guidelines and form, however, as is the case in most multi-center studies no control could be exerted over the local site EHRs or clinical documentation practices. With respect to the training, the standardized cases used were created from our EHR and thus, the accuracy measurement may not have been representative of accuracy of abstraction from the local site's EHR, i.e., a threat to the validity of the accuracy measurement. Threats to generalizability include our results representing medical record abstraction training for only one study in one clinical area. The percentage of study data available in EHRs varies by institution and clinical specialty and individual study as do the location of data within the EHR and the processes by which they arrive there. However, the challenges and lessons learned in developing and delivering medical record abstraction training in the context of a multi-center study are not dependent on the study or clinical specialty.

7. Conclusions

Given the twenty-three percent of trainees exceeding the acceptance criteria, the training and quantification of the error rate was deemed necessary to assure that data are capable of supporting study conclusions. Though few studies report active quality control of medical record abstraction or associated discrepancy rates, given the results of the previously reported pooled analysis and our own early training experience, i.e., the magnitude and variability of the error rate, we deemed the planned ongoing independent re-abstraction and measurement of inter-rater reliability on a sample of cases throughout the study to be necessary. Based on the literature and our experience here, we recommend undertaking MRA training for multi-center studies that rely on data abstracted from medical records.

References

1. M.N. Zozus, et al., *Factors Affecting Accuracy of Data Abstracted from Medical Records. PLoS One,* **10**(10) (2015), e0138649.
2. M. Nahm, et al. *Distributed Cognition Artifacts on Clinical Research Data Collection Forms*. in *AMIA CRI Summit Proceedings*. 2010. San Francisco, CA.
3. M. Nahm, *Data quality in clinical research*, in *Clinical research informatics*, R. Richesson and J. Andrews, Editors. 2012, Springer-Verlag: New York.
4. *Good Clinical Practice: Integrated Addendum to ICH E6(R2) Guidance for Industry*, U.S.D.o.H.a.H. Services, Editor. March 2018, U.S. Food and Drug Administration
5. Food and Drug Administration, U.S.D.o.H.a.H.S., *Use of electronic health record data in clinical investigations guidance for industry*, U.S.D.o.H.a.H. Services, Editor. 2018, Food and Drug Administration.
6. K.M. Newton, et al., *Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Inform Assoc*, **20**(e1) (2013), e147-54.
7. J. Snowden, et al., *The institutional development award states pediatric clinical trials network: building research capacity among the rural and medically underserved. Curr Opin Pediatr*, **30**(2) (2018), 297-302.
8. Gibbs, D., *For Debate: 250th anniversary of source document verification. BMJ*, 1996. **313**.
9. Feinstein, A.R., J.A. Pritchett, and C.R. Schimpff, *The epidemiology of cancer therapy. 3. The management of imperfect data. Arch Intern Med,* **123**(4) (1969), 448-461.
10. Feinstein, A.R., J.A. Pritchett, and C.R. Schimpff, *The epidemiology of cancer therapy. IV. The extraction of data from medical records. Arch Intern Med,* **123**(5) (1969), 571-590.
11. Burnum, J.F., *The misinformation era: the fall of the medical record. Ann Intern Med*, **110**(6) (1989), 482-484.
12. Lei, J.v.d., *Use and abuse of computer-stored medical records. Meth Inform Med*, **30** (1991), 79-80.
13. Meads, S. and J.P. Cooney, *The medical record as a data source: use and abuse. Top Health Rec Manage*, **2**(4) (1982), 23-32.
14. Hogan, W.R. and M.M. Wagner, *Accuracy of data in computer-based patient records. J Am Med Inform Assoc,* **4**(5) (1997), 342-355.
15. Thiru, K., A. Hassey, and F. Sullivan, *Systematic review of scope and quality of electronic patient record data in primary care. BMJ*, **326**(7398) (2003) 1070.
16. Chan, K.S., J.B. Fowles, and J.P. Weiner, *Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev,* **67**(5) (2010), 503-527.
17. United States Department of Health and Human Services, F.a.D.A., *Guidance for industry oversight of clinical investigations — a risk-based approach to monitoring* U.S.D.o.H.a.H. Services, Editor. 2013, Food and drug administration.
18. Nahm, M., et al., *Data Quality Survey Results. Data Basics,* **10** (2004), 7.
19. Rostami, R., M. Nahm, and C.F. Pieper, *What can we learn from a decade of database audits? The Duke Clinical Research Institute experience, 1997--2006. Clin Trials*, **6**(2) (2009), 141-150.
20. Management, S.f.C.D., *Good Clinical Data Management Practices (GCDMP)*. 2013, Society for Clinical Data Management.

Appendix: Operational lessons Learned

1. We initially wanted training cases from three institutions so site-to-site variability would be reflected in the standardized training cases. One of the institutions did not have the time to submit to their IRB for approval for secondary use of data to create the cases or to create the training cases. Two institutions created de-identified, standardized mock cases for training. Training cases were provided hard-copy and from the two institution's EHRs. rather than training on cases from multiple EHRs.

2. Study coordinators at the coordinating center were used to create the standardized training cases. Directions to the study coordinator were not sufficiently clear and suffered from the following problems: (1) The initial cases consisted of printed "evidence" to support the completed data collection form rather than a representation of the full chart from the NICU stay. (2) In remediating the first issue, because the cases had been redacted, it was not possible to locate the initial example in the EHR and the case was filled-in with pages from other patients. This caused inconsistencies in gender and care progression across the initial two cases that required further correction. (3) Items were missed in redaction despite two independent reviews of redaction. Because of quality issues with consistency and redaction of the cases, additional redaction work on the cases was required.

3. The initial training plan was designed based on sites abstracting two cases from their institutional EHRs and following the study abstraction guidelines prior to the training. The purpose was that trainees would come to the training sessions with questions. Because sites were in start-up at the time of the training, most training occurred prior to IRB approval for each site. Thus, sites were not able to do the pre-training assignment and entered the training, in some cases, without having attempted to use the abstraction guidelines.

4. There were significant concerns from the study team that the training would delay study start-up. More time upfront and better planning would have allowed more notice to sites and prevented site start-up delays due to the training. Providing time on task estimates to trainees/sites in advance would have also helped sites plan time for the training.

5. In the training, we encountered several cases where abstractors who care for NOWS babies regularly had difficulty with adhering to the abstraction guidelines. Based on discussions with the abstractors and clinical operations at the coordinating center, we have come to believe that the abstractors relied on their clinical knowledge rather than the abstraction guidelines. We note that the 2015 review and Delphi process also uncovered the debate over advantages and disadvantages of using clinical abstractors.[1] Based on our experience to date, we as well remain undecided on the issue. The solution proposed by one study abstractor was to state explict operationalizations as definitions of key data items, such as "rooming in" or other non-pharmacological interventions.

6. Abstraction trainees were provided their error rate and general "common mistakes and problems". Due to tight start-up timelines and concerns about the abstraction

training delaying study start-up, most abstractors did not receive individualized feedback on their specific errors.

7. Sites distributed abstraction tasks differently. PIs at some sites took the training with the intent of serving as the independent abstractor for the study. Most sites provided study coordinators and research nurses, while others sent research assistants who enter data to the abstraction training. A statement of the intended and target audience, as well as pre-requisite knowledge such as knowledge of the local EHR or chart at the site and knowledge of local clinical documentation practices, would have helped sites select abstractors. Though this would not alleviate the variability of personnel skills or the need to assign available people where availability of experienced abstractors is limited.

8. The abstraction training revealed gaps in the abstraction guidelines. The gaps were provided to the study PIs who chose which to incorporate. In all, nine changes to the abstraction guidelines were suggested based on the training experience. We include these and others that have arisen since.

   a) Both test cases were deliveries of a caucasian female with no mention of ethnicity. Some sites selected "Not Hispanic or Latino" vs. "Unknown or Not Reported". We emphasized in the abstraction guidelines that if the medical record did not document ethnicity, it should be recorded on the data collection form as "Unknown or Not reported".

   b) A related situation occurred regarding the baby's race. Multiple sites recorded the mother's race where the baby's race was not found in the record. The data collection form was updated to only require the mother's race.

   c) The data collection form asks sites to select the location/nursey type: (Level 1,2,3,4 or pediatric unit). The most common deficiency was that sites did not select all locations where the infant(s) received care. The abstraction guidelines were updated to emphasize (bolded and underlined) "ALL".

   d) The data collection form asks if there were any diagnoses recorded in the medical record that may have contributed to a lengthened hospital stay other than NOWS, and the most common non-NOWS factors selected are: social, hyperbilirubinemia, and respiratory illness. Hyperbilirubinemia and/or respiratory illness may have initially been a problem list item, however, was resolved early during the subject's hospital stay. The most recent guidelines had already clarified how to handle this situation.

   e) The maternal section of the data collection form asks if the prenatal care was adequate, inadequate, none, or unknown. The EHR documentation for training case 001, stated that prenatal care was limited. Many trainees selected "Inadequate" vs. "unknown". The MOP defines inadequate prenatal care as less than 3 visits or prenatal care started in the 3rd trimester. The abstraction guidelines at the time stated that, "if the only information was the word limited, unknown should be selected". No further updates were made.

   f) The non-pharmacologic section of the data collection form requires selection of the non-pharmacologic support the infant received from the following list: Rooming-in, Clustered care, Swaddling, Kangaroo Care/Skin-to-Skin, Low Lights, reduced noise, Parental Education, Non-Nutritive Sucking/Pacifier. …. Some abstractors left this section blank or indicated minimal support. These items are typically found in the flowsheet assessments and progress notes/record. The abstraction guidelines and training were updated to indicate the values from the flow sheets as the primary source and notes secondary.

g) The Scoring/Assessment System section of the data collection form requires indication of the assessment system. Both training cases did not state which assessment tool was used, however, scores were provided. Because this is generally a facility-level decision, we assume that hospitals only use one scoring system and know the system. Therefore, where not stated in the EHR, the data collection form may serve as the source docment for this data. We assume that facilities can produce documentation of an institutional decision to use a particular scoring system.

h) The data collection form asks, "Where was the infant discharged or transferred to" and uses the following response list: home with parent, home with relative, home with foster parent, transferred to another hospital, or transferred to outpatient treatment center. For training case 2, it was not documented in the test EHR data with whom and to where the infant was discharged. Most abstractors indicated that the infant was discharged with the parent. Unknown was not an option for version 3.0 of the data collection form. The data collection form was updated to replace "home with foster parent" as a response option with "home with foster/adoptive parent". "Other, specify" was added as an option for this question in the subsequent version of the data collection form.Abstraction training and guidelines were modified and updated training emphasized referencing the abstraction guidelines.

i) In the same section, the data collection form asks whether outpatient follow-up was planned/scheduled prior to discharge. For training case 001, the discharge follow up was scheduled with Dr. [REDACTED] at the [REDACTED neighborhood name] Clinic. Most abstractors selected, "Primary Care Physician" vs. "Neonatal Follow up Clinic or "Other, specify". The abstraction guifelines and training were updated to emphasize that information not explicitly stated in the EHR should be reported as unknown.

j) A post-training question arose regarding classification of infants born in an ambulance enroute to the hospital. The abstraction guidelines define infants born in an ambulance as outborn.