OPEN ACCESS

Phenotyping for patient safety: algorithm development for electronic health record based automated adverse event and medical error detection in neonatal intensive care

Qi Li,[1] Kristin Melton,[2] Todd Lingren,[1] Eric S Kirkendall,[1,3,4] Eric Hall,[2] Haijun Zhai,[1] Yizhao Ni,[1] Megan Kaiser,[1] Laura Stoutenborough,[1] Imre Solti[1,4]

[1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[2]Division of Neonatology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[3]Division of Hospital Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA
[4]James M Anderson Center for Health Systems Excellence, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA

Correspondence to
Dr Imre Solti, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, 3333 Burnet Avenue, MLC 7024, Cincinnati, OH 45229-3039, USA;
imre.solti@cchmc.org

Open Access
Scan to access more
free content

CrossMark

ABSTRACT

Background Although electronic health records (EHRs) have the potential to provide a foundation for quality and safety algorithms, few studies have measured their impact on automated adverse event (AE) and medical error (ME) detection within the neonatal intensive care unit (NICU) environment.

Objective This paper presents two phenotyping AE and ME detection algorithms (ie, IV infiltrations, narcotic medication oversedation and dosing errors) and describes manual annotation of airway management and medication/fluid AEs from NICU EHRs.

Methods From 753 NICU patient EHRs from 2011, we developed two automatic AE/ME detection algorithms, and manually annotated 11 classes of AEs in 3263 clinical notes. Performance of the automatic AE/ME detection algorithms was compared to trigger tool and voluntary incident reporting results. AEs in clinical notes were double annotated and consensus achieved under neonatologist supervision. Sensitivity, positive predictive value (PPV), and specificity are reported.

Results Twelve severe IV infiltrates were detected. The algorithm identified one more infiltrate than the trigger tool and eight more than incident reporting. One narcotic oversedation was detected demonstrating 100% agreement with the trigger tool. Additionally, 17 narcotic medication MEs were detected, an increase of 16 cases over voluntary incident reporting.

Conclusions Automated AE/ME detection algorithms provide higher sensitivity and PPV than currently used trigger tools or voluntary incident-reporting systems, including identification of potential dosing and frequency errors that current methods are unequipped to detect.

OBJECTIVE

We present the findings of our work to identify adverse events (AEs) or medical errors (MEs) using the electronic health record (EHR) in the neonatal intensive care unit (NICU) of one of the largest tertiary pediatric hospitals in the USA.[1] Our goal is to develop novel algorithms for automated AE/ME detection, and compare the performance of these algorithms to the current error detection systems, including trigger tool and voluntary incident-reporting systems. We describe the study design, including the data sources being used for algorithm development and a description of the annotation techniques used to build the natural language processing (NLP) component of subsequent algorithms. We also present the findings of two AE/

ME phenotyping algorithms (IV infiltrates and narcotic medication events) that leverage structured EHR content.

BACKGROUND AND SIGNIFICANCE

The federal government is on track to spend up to $27 billion in incentives for physicians and hospitals to install EHRs.[2] Although the planned widespread implementation of EHRs brings the promise of abundant resources for research purposes via secondary uses of EHR data,[3] there are few studies that measure how well the EHR contributes to the automated detection of AEs and harms in the neonatal population. Although several large-scale studies have focused on the Institute for Healthcare Improvement (IHI)'s Global Trigger Tool,[4][5] a recent literature review found a lack of systematic studies that compare multiple methods for error and harm detection to each other and to a scientifically generated measurement standard.[6]

Many gaps remain in our understanding of AEs/MEs that occur in the NICU.[7] Trigger tools and voluntary incident-reporting systems are the most prevalent error-detection methods in healthcare institutions.[8–10] Other methodologies include screening administrative data (eg, ICD-9-CM) codes.[11–13] However, the raw data that are collected in the EHR during routine clinical care provide a potentially more granular, precise, and comprehensive view of patients and the thought processes of physicians. Furthermore, increasing evidence suggests that EHR data have the potential to provide a foundation for quality and safety algorithms. A 2003 study showed that 66% of indicators needed to assess the RAND Quality Assessment Tools' 37 clinical areas were absent from claims data (367 of 553 indicators).[13] In addition, approximately 50% of the quality indicators require information that is available only in clinical notes. Murff *et al* concluded that NLP-based analysis of EHRs proved superior to traditional patient-safety indicators and administrative code-based algorithms at identifying postoperative complications, while Melton and Hripcsak recently showed the utility of NLP-based AE detection from discharge summaries.[14][15]

Definitions

An AE is generally defined as 'injuries, large or small, caused by medical management rather than the underlying condition of the patient'. An ME is defined as 'the failure of a planned action to be

completed as intended or the use of a wrong plan to achieve an aim'.[16] [17] The distinction is important because many MEs do not lead to harm, while AEs, by definition, are linked to patient harm but *may not* be due to an ME.[18] Medication/fluid errors are the most frequently described safety issues in the neonatal population, while airway management AEs are frequently associated with harm.[19] We have categorized our results as AEs or MEs, but the algorithms are designed to detect both types of events. To specify the combined detection by the algorithms throughout the paper, AE/ME will refer to 'adverse event or medical error'.

In the context of the current paper, phenotyping is defined as EHR-based algorithmic identification of cohorts of patients with AE/ME. Annotation is defined as manual identification of AE/ME in NICU notes using computational linguistics development practices.

DATA AND METHODS

We compiled a comprehensive list of AE/ME for four selected categories that occurred between January 1, 2011 and December 31, 2011 in the Cincinnati Children's Hospital Medical Center (CCHMC) NICU: IV infiltrates, medication and fluid errors, airway management AEs, and code AEs/MEs. These AEs/MEs are as shown in figure 1.

Population, environment, and EHR

CCHMC is a large, urban pediatric academic medical center. The CCHMC NICU is a level IV NICU providing extracorporeal membrane oxygenation (ECMO), surgical, and subspecialty care to neonates with complex medical needs. The majority of patients are outborn, with an average gestational age of 35 weeks and an average length of stay of 26.6 days. The unit, staffed by 11 faculty physicians, has an average daily census of 45 patients.

An industry-leading EHR with an order entry system (CPOE) and electronic documentation (Epic Systems Corporation, Verona, Wisconsin, USA) has been fully implemented since 2010. CCHMC also has a robust history in AE/ME detection, utilizing multiple modalities such electronic reporting, manual chart review, and voluntary incident-reporting methodologies.[9] [20] There are two layers of decision support for narcotic administration in the CPOE system. One layer provides standard weight-based doses to guide the user to correct dosing as the order is written. The second layer provides assessment of the dosing range at order signing, with an alert for overdoses that can be overridden.

Trigger tool system dataset

One method currently used for AE detection is a trigger tool methodology. CCHMC's Automated Adverse Event Detection Program (AAEDP) was established in 2006 and now includes over a dozen electronic safety-surveillance triggers. Triggers are data or combinations of data that may signal an underlying event of interest. Triggers are activated in near-real time and automatically populate an electronic data repository and reporting database. The database has a web-based interface that expert nurse reviewers and safety-subject-matter experts use to view,
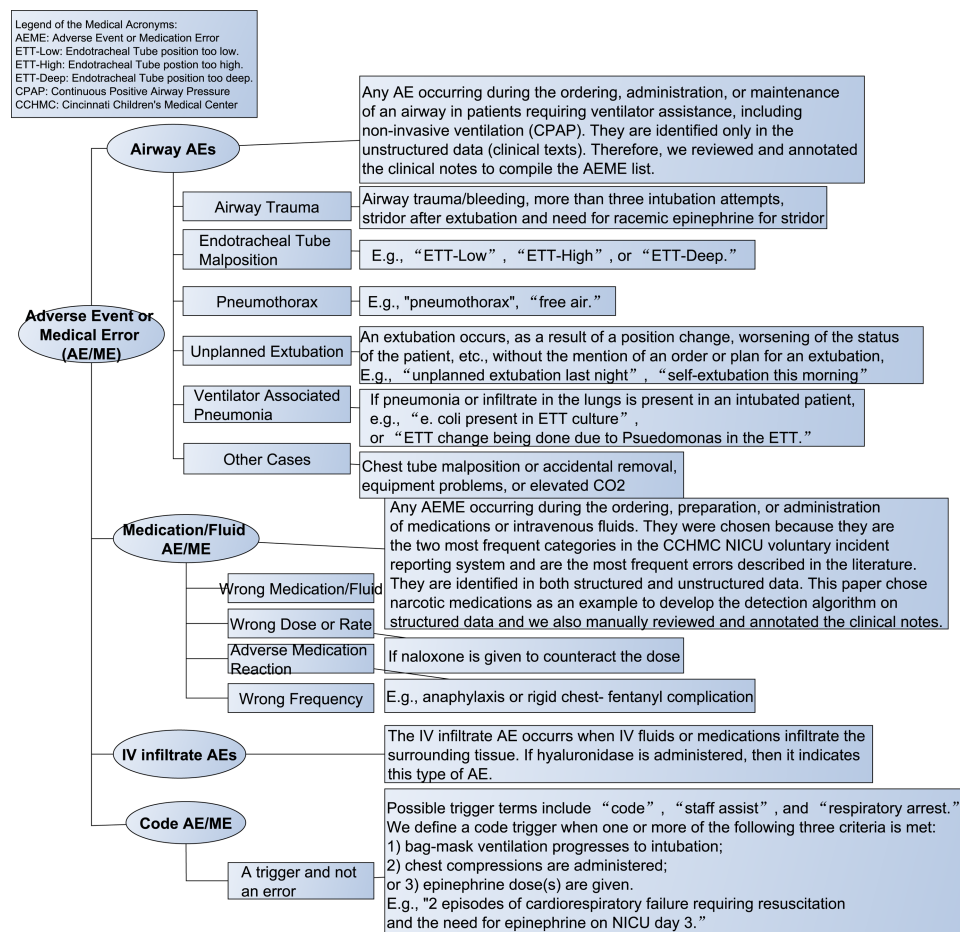

Figure 1 Adverse event (AE)/medical error (ME) definitions and descriptions.

manually validate, and adjudicate EHR data for possible AE/ME signals. Triggers evaluated in this study for comparison to AE/ME detection algorithms include the trigger detection of hyaluronidase administration and naloxone administration.

Voluntary incident reporting dataset

Incident reports are another AE/ME detection strategy. Our institute uses Risk MonitorPro to collect voluntary incident reporting.[21] Reports can be submitted either anonymously or with identification, and by any hospital employee. Access to the program is provided on all hospital computers through either an intranet link or directly through an EHR user interface. A structured online reporting form is utilized, and includes: (1) patient information, (2) date and location of the error or near-miss, (3) general incident categories, (4) factors that contributed to the event, (5) actions to mitigate the event, and (6) assessment of patient harm. The form also includes an area to provide a brief factual description of the event using free text and an area for feedback after review.

We analyzed 2011 reports submitted through the voluntary incident-reporting system. Excluding laboratory specimen reports, 350 NICU events were reported during the study period. In our areas of interest, 79 medication/fluid errors and 26 airway management issues were reported. Individual incidents were manually reviewed and the analysis identified one narcotic medication bolus dosing error, one narcotic frequency error, 20 unplanned extubation events, and four IV infiltrates that required treatment.

EHR dataset

Under an approved IRB protocol, data from the EHR, trigger tool system, and voluntary incident-reporting system were collected from all 753 CCHMC NICU patients in 2011, which represented a total of 20 140 patient days.

This study collected the EHRs for each NICU patient day of our included cohort, including the structured data of the medication order history (Order), audit trail (Audit), medication administration record (MAR), and lab results (Lab), as well as the unstructured data of note documentation, as described in table 1.

AE/ME detection methods

Phenotyping for patient safety in NICU patients required mapping data across multiple sources (EHR content, trigger tool database, and voluntary incident reporting) and formats (structured and unstructured data). Figure 2 depicts the structure of the AE/ME detection methods using EHR structured and unstructured data, and evaluation results compared with trigger tool and voluntary incident-reporting systems as well as manual chart reviews validated by an NICU physician.

IV infiltration AE detection algorithm

The IV infiltration algorithm (figure 3A) monitors the structured data for use of the trigger medication hyaluronidase to determine if any severe IV infiltrate occurs. Clinically, the only indication for hyaluronidase use in the NICU is for treatment of severe IV infiltrates. It has no other clinical use, and thus can be used to identify the occurrence of severe infiltrates. This trigger has also been implemented by manual ME trigger tools.[7] Less severe infiltrates not requiring hyaluronidase may be described in the clinical notes. Therefore, we also annotated clinical notes for IV infiltrates.

Narcotic medication AE/ME detection algorithm

The narcotic medication AE/ME detection algorithm is shown in figure 3B. As narcotic medications are often used in the NICU, naloxone is administered to reverse narcotic effects; it is generally given when a patient has had an AE with narcotic administration, such as apnea or respiratory depression. Therefore, administration of naloxone represents an AE trigger for narcotic oversedation. The trigger or unexpected sign and symptom step in the algorithm mirrors the approach that has been implemented by widely-used manual ME trigger tools.[7] In addition to narcotic oversedations, non-standard narcotic orders, narcotic medication delivery errors, and frequency errors were also detected.

Detection of non-standard dosing

To assess periods where patients were at higher risk for narcotic dosing errors, the frequency with which narcotic medications were ordered at doses that exceeded standard dosing was evaluated. This method does not detect errors per se, but detects vulnerability to error based on the use of non-standard dosing. We used published recommendations and local clinical practice guidelines as standards for morphine and fentanyl dosing, as shown in table 3.

We also assessed dosing escalation for continuous narcotic medications to determine the number of times doses are

Table 1 Description and descriptive statistics of the NICU 2011 EHR data

	Notes	Medication order	Audit	MAR	Lab
EHR data description	Clinical notes include procedure notes (eg, intubation notes) and progress notes (eg, surgical progress note, attending notes, postoperative notes, plan of care notes, nursing notes and consult notes)	Medication order describes medications ordered by the physicians, including: ▶ medical record number ▶ medication order ID ▶ medication name ▶ ordered dose strength ▶ ordered dose frequency ▶ medication start date ▶ medication end date ▶ dosing weight	Audit data describes the medication order changes by the physicians, including the dosing changes	MAR is a record of the administered medications, including: ▶ actual dosage ▶ strength ▶ route ▶ time of administration for ordered medications	Lab results provide numerical values and units for each test, and report the time the specimen was obtained (eg, blood drawn)
Number of study-specific unique entries/objects	30 115	38 282	405 519	180 595	333 014

EHR, electronic health record; MAR, medication administration record; NICU, neonatal intensive care unit.
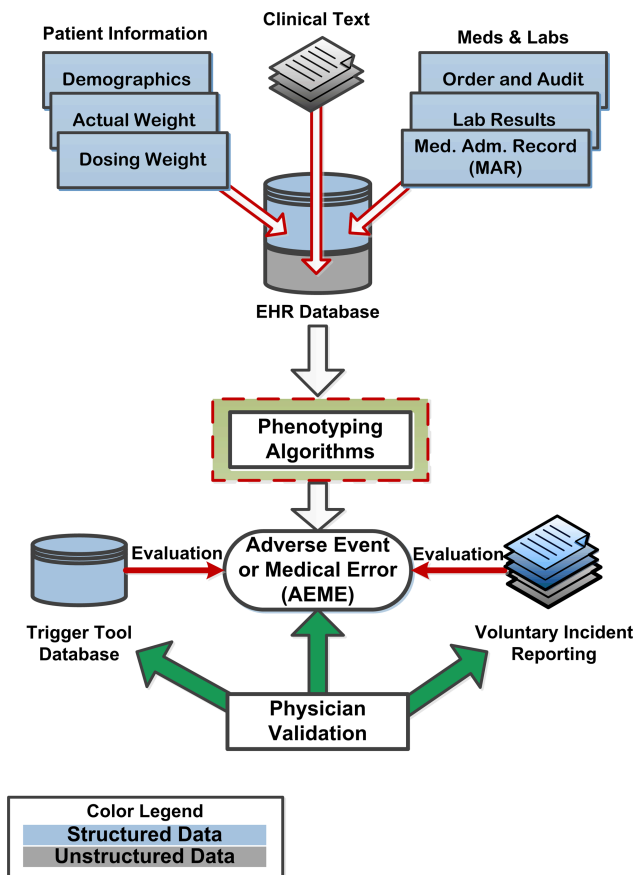
Figure 2   Overall adverse event (AE)/ME detection methods and evaluation method with main data sources and formats.

escalated more rapidly than standard management. Escalation that exceeds the standard dose is not an error itself, but represents a clinical practice variation that increases vulnerability to error. To perform this assessment, we assumed standard dosing increases of ≤0.1 mg/kg/h for continuous morphine, and ≤1 μg/kg/h for continuous fentanyl.

Medication delivery error detection

This evaluates for agreement between the medication order and the medication delivery, using the medication order history, order audit trail, and MAR data. Morphine and fentanyl bolus doses were evaluated for agreement.

Assessment of medication frequency errors

We evaluated potential frequency errors for narcotic bolus medications ordered as scheduled doses (SCHEDULED) or 'as needed' doses (PRN). The SCHEDULED category has the frequency description of 'every X hours/days/etc', such as 'every 1 h', which theoretically requires an exact delivered time. It is difficult, however, to deliver a medication exactly within a specified time window. Therefore, we evaluated the number of events that fell outside of the specified window, and report the results according to the time discrepancy in delivery. For example, if a medication is ordered every 1 h and the medication is delivered at exactly 1 h, we report a 0-discrepancy. Subsequently, 10% of 1 h ordered frequency allows a 6 min time discrepancy; 20% allows a 12 min time discrepancy; and so on. Different order frequencies (ie, every 4 h) result in different overall time discrepancies that are allowed, but are characterized as the same percentage of the ordered frequency. Any cases where the medication is administered beyond the allowed discrepancy window are reported as out-of-boundary cases. The PRN category has the frequency description of 'every X hours as needed', which means the medication delivered time cannot be earlier than the X-hour window.

Phenotyping AE/ME from unstructured data

AE/ME annotation in the clinical text is required for EHR-based machine-learning or rule-based detection from unstructured data as a gold standard. In this paper, we report our manual AE/ME annotation from clinical notes. All notes were double annotated by two annotators and were adjudicated under the supervision of an NICU physician and the annotation manager. Discrepancies were resolved by the NICU physician. Both annotators are English speakers, with at least 1 year of clinical text annotation experience, with bachelor's degrees (one a clinical
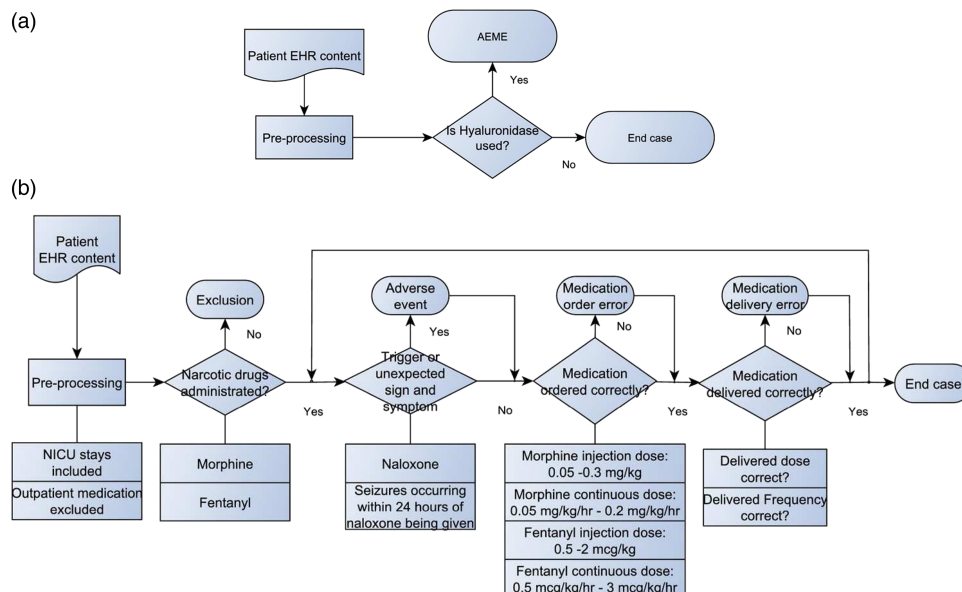

Figure 3   (A) The neonatal intensive care unit (NICU) electronic health record (EHR)-based phenotyping adverse event (AE) detection algorithm: IV Infiltration. (B) The NICU EHR-based phenotyping adverse event (AE)/medical error (ME) detection algorithm: narcotic drugs.

Table 2 Severe IV infiltrate, narcotic oversedation, and narcotic bolus adverse event (AE)/medical error (ME) detection

Gold standard		Detected by algorithm	Detected by trigger tools	Detected by incident reporting
Severe IV infiltrate AEs (ie, hyaluronidase treatment)				
12	PPV	100%	91.7%	100%
	Sensitivity	100%	91.7%	33.3%
	Specificity	100%	99%	100%
	Detected	12	12	4
	True error	12	11	4
	False positive	0	1	0
Narcotic oversedation AEs (ie, naloxone treatment)				
1	PPV	100%	100%	0%
	Sensitivity	100%	100%	0%
	Specificity	100%	100%	100%
	Detected	1	1	0
	True error	1	1	0
	False positive	0	0	0
Narcotic bolus MEs				
Morphine				
10 (out of 5641)	PPV	43.5%*	0%	100%
	Sensitivity	100%	0%	10%
	Specificity	99.8%	100%	100%
	Detected	23	0	1
	True error	10	0	1
	False positive	13, including 5 small amount discrepancy and 8 documentation errors	0	0
Fentanyl				
7 (out of 1464)	PPV	38.9%†	0%	0%
	Sensitivity	100%	0%	0%
	Specificity	99.2%	100%	100%
	Detected	18	0	0
	True error	7	0	0
	False positive	11, including 5 small amount discrepancy and 8 documentation errors	0	0

*If considering the documentation errors as true errors, then our algorithm has PPV of 78.3%.
†If considering the documentation errors as true errors, then our algorithm has PPV of 66.7%.
PPV, positive predictive value or precision.

Table 3 Non-standard narcotic dosing detection

	Standard dose*	Total orders	Out of boundary cases	Out of boundary rate (%)
Out-of-boundary standard dose detection				
Morphine				
Continuous (mg/kg/h)	0.05–**0.2**	300	38	12.7
Injection (mg/kg)	0.05–0.1 (usual dose); 0.05–**0.3** (acceptable dose)	1559	27	1.7
Fentanyl				
Continuous(μg/kg/h)	0.5–**3**	85	28	32.9
Injection (μg/kg)	0.5–2 (usual dose); 2–**5** (acceptable dose)	603	6	1.0

Starting dose	Increasing dose	Total orders	Out of boundary cases	Out of boundary rate (%)
Out-of-boundary continuous dosing detection with the normal escalation estimation				
Morphine (mg/kg/h)				
0.1	0.1	300	33	11.0
0.2	0.1		16	5.3
0.3	0.1		5	1.7
0.4	0.1		5	1.7
0.5	0.1		1	0.3
Fentanyl (μg/kg/h)				
1	1	85	18	21.2
2	1		13	14.3
3	1		9	10.6
4	1		5	5.9
5	1		3	3.5

*Bold type indicates the upper bound.

RN and one a BSN degree). We used the Knowtator plug-in for Protégé for the annotation task.[22] To define the annotation guidelines, we derived 11 classes for medical code (one class), airway (six classes), and medication/fluid (four classes) AE/MEs, as shown in figure 1.

During the annotation training period, we iteratively developed and refined annotation guidelines through consultation with an NICU physician. We designated the first period of annotation as 'pre-training' because the annotation guidelines had not been finalized; in this period, we annotated the procedure notes as a potentially rich source of errors. The second period of annotation was designated 'training-1'. Twenty out of the 753 patients had been identified as having unplanned extubations in voluntary incident reporting, and another 11 cases were identified via other means in NICU data. We selected the progress notes for these patients to annotate for two reasons: (1) because annotating known errors would prove useful training material, and (2) to develop a comparison set against the voluntary reporting tool. The third period, or 'training-2', contains the annotation of the final 590 notes in the patient-specific group of progress notes. We report the inter-annotator agreement (IAA) using the F-measure and the final AE/ME instances in each category.

Algorithm development and AE/ME detection validation by chart review
Algorithm rules were specified by the neonatologists and implemented by the programmer. We developed and evaluated the algorithms in the following research workflow:

1. The neonatologists provided algorithm specifications in a written document and in an initial discussion.
2. Based on the specifications, the programmer coded the algorithm.
3. On a mock dataset the programmer checked the algorithm for compliance with the physician's specifications. This third step was an engineering quality assurance step and its sole purpose was to ascertain that the algorithms were implementing the specifications of the physicians.
4. The algorithm was executed on the EHR data.
5. The neonatologists manually checked all medication lists for the 753 NICU patients for IV infiltrates and narcotic oversedations. All narcotic medication orders and administrations from 328 patients receiving narcotics (morphine and fentanyl) were also reviewed for narcotic delivery errors and the findings were compared to the algorithm output. In total 7118 medication events were manually checked. We report the positive predictive value, sensitivity, and specificity of the algorithms. The neonatologist reviewed the narcotic non-standard dosing and time discrepancy cases. The same expert similarly validated the adjudicated results of the clinical note annotation. The trigger tool and voluntary incident-reporting data sets were validated by physician experts at the time of their collection and revalidated by the neonatologist.

RESULTS
IV infiltrate AE detection algorithm
The IV infiltrate AE algorithm detected 12 hyaluronidase treatments for severe IV infiltrates that occurred in 2011, and all of them were confirmed by the gold standard as true AEs. However, only 11 infiltrates were detected by the trigger-tool methodology with one false positive detected error. Four infiltrates were identified by the incident-reporting system (table 2).

Narcotic medication AE/ME detection algorithm
The narcotic medication AE/ME algorithm detected one naloxone treatment in the NICU for narcotic oversedation, which represented 100% sensitivity and agreement with the trigger tool. In the clinical notes, we did not find seizures occurring during the 24 h period after naloxone was given.

The second part of the algorithm evaluates whether delivered narcotic bolus doses match the ordered doses. The algorithm detected 10 morphine delivery errors (0.18%, 10/5641) and seven fentanyl delivery errors (0.48%, 7/1464), where delivered doses were higher or lower than the ordered doses, as true errors. The algorithm also detected five morphine and six fentanyl small-amount discrepant doses (<0.05 mg for morphine and <0.1 μg for fentanyl), suggesting that discrepancies occurred based on the use of actual patient weight versus dosing weight. Finally, the algorithm identified documentation errors for morphine (n=8) and fentanyl (n=5) where the documented doses reflected the use of a full medication vial and not the actual delivered dose, making error detection unreliable. The small discrepancies and the documentation errors were counted as false positive cases. In comparison, medication delivery errors are not detected by trigger tool assessments, and only one narcotic bolus delivery error was reported by incident reporting, representing 16 more narcotic delivery errors detected by the AE/ME algorithm than by the incident-reporting system (table 2).
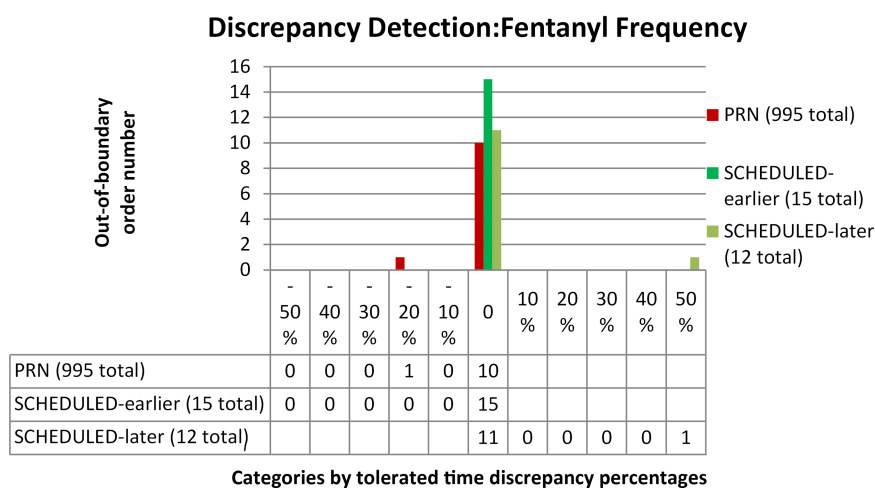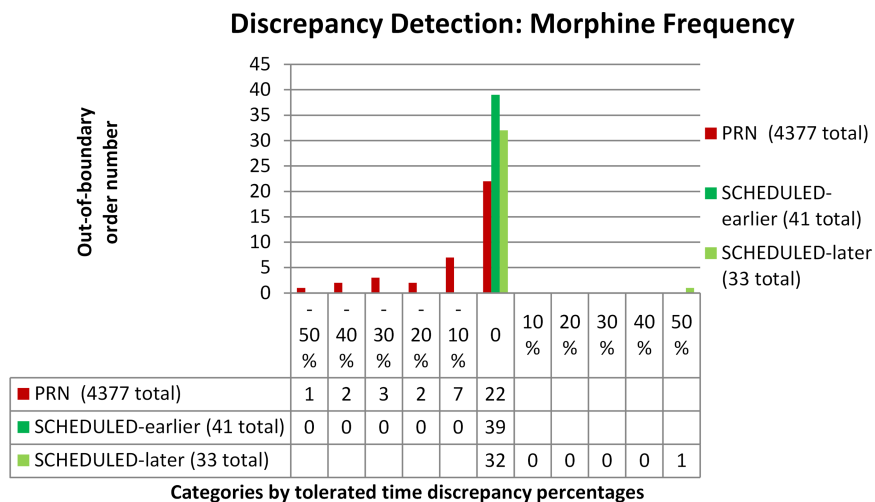
Non-standard narcotic dosing
We evaluated the frequency with which ordered narcotic doses exceed standard dosing as a means to assess vulnerability to error. Using published and local dosing standards, our algorithm found that 12.7% (38/300) of continuous morphine orders, 1.7% (27/1559) of morphine injection orders, 32.9% (28/85) of continuous fentanyl orders, and 1.0% (6/603) of fentanyl injection orders were outside of standard dosing ranges, as shown in table 3. We also evaluated narcotic dosing escalation as a way to look for continuous narcotic orders that may be more prone to error according to standard clinical practice. We report the results with morphine escalated by 0.1 mg/kg/h from any starting dose, and fentanyl escalated by 1 μg/kg/h from any starting dose. Orders that escalated doses by larger amounts are reported as an out-of-boundary order and are shown in table 3. For example, 11% of continuous morphine doses that were started at 0.1 mg/kg/h were escalated by more than 0.1 mg/kg/h.

Frequency error detection
We report the frequency errors according to the frequency discrepancy percentage in three categories: PRN and SCHEDULED, including early and late doses. Figure 4 shows the number of narcotic bolus deliveries that fall into the different discrepancy time windows. If we set the tolerated frequency discrepancy to zero (0% time discrepancy), then the number of discrepant doses are 37 (37/4377) for the morphine PRN category, 72 (72/74) for the morphine SCHEDULED category, 11 (11/995) for the fentanyl PRN category, and 27 (27/27) for the fentanyl SCHEDULED category. If we set the tolerated frequency discrepancy to 10%, then the number of discrepant doses decreases to 15, 1, 1, and 1 cases for the morphine PRN, morphine SCHEDULED, fentanyl PRN, and fentanyl SCHEDULED categories respectively. As shown in figure 4, the majority of doses are delivered with only a small discrepancy (<10%) from their ordered time. However, the algorithm identified multiple PRN doses that were given early (20–50% discrepancy) without an order and also identified one

Figure 4 Narcotic bolus frequency discrepancy detection.

Discrepancy Detection: Morphine Frequency

	- 50 %	- 40 %	- 30 %	- 20 %	- 10 %	0	10 %	20 %	30 %	40 %	50 %
■ PRN (4377 total)	1	2	3	2	7	22					
■ SCHEDULED-earlier (41 total)	0	0	0	0	0	39					
■ SCHEDULED-later (33 total)						32	0	0	0	0	1

Categories by tolerated time discrepancy percentages

Discrepancy Detection:Fentanyl Frequency

	- 50 %	- 40 %	- 30 %	- 20 %	- 10 %	0	10 %	20 %	30 %	40 %	50 %
PRN (995 total)	0	0	0	1	0	10					
SCHEDULED-earlier (15 total)	0	0	0	0	0	15					
SCHEDULED-later (12 total)						11	0	0	0	0	1

Categories by tolerated time discrepancy percentages

SCHEDULED dose each of morphine and fentanyl that was given with a 50% discrepancy from their scheduled time. In comparison, incident reporting identified just one scheduled morphine dose ordered every 4 h and delivered at 2 h, consistent with a 50% time discrepancy.

Results of phenotyping AE/ME annotation
Given that the bulk of AE/ME documentation is contained in the unstructured data of the EHR, we annotated clinical notes to detect medication, airway event, and code event AE/MEs. Table 4 shows descriptive, IAA statistics (F-measure) and the AE/ME instance numbers of the two types of notes annotated during the pre-training and training periods. A higher proportion of procedure notes contained AE/MEs than progress notes (11.6% vs 5.3% and 10.3%, table 4). We report the IAA for AE/MEs that had at least 20 documented instances in the three annotation periods. We detected improvement in IAA in the training period between procedure notes (pre-training) and progress notes (training). For endotracheal tube (ETT) malposition, unplanned extubation, and other classes, the greatest IAA was seen in the training-2 period (100%, 96%, and 72.7%, respectively). IAA for airway trauma, pneumothorax, and code event increased from the pre-training to the training-1 period, but dropped in the training-2 period.

Overall we detected 420 AE/ME instances in 3263 clinical notes. The data represent the AE/ME detection in the reviewed notes, but one AE/ME may be represented multiple times, either in the same notes or in several notes and will require post-processing after the completion of annotation to determine unique AE/ME frequency. Airway AE/MEs were more frequently described in the clinical notes than medication/fluid AE/MEs. Several event instances (airway trauma, ETT malposition, pneumothorax, other) occurred more frequently in procedure notes. Code event instances also occurred more frequently in procedure notes. Medication-related AE/ME instances were very infrequent in both note types.

DISCUSSION
Phenotyping AE/ME detection algorithm from structured data
This work developed and evaluated two phenotyping AE/ME detection algorithms for severe IV infiltration and narcotic related AE/ME. The novel algorithms were compared to the events identified by trigger tools and voluntary incident-reporting systems. Although incident reporting is a widely used method for identifying AE/ME, it is prone to underreporting and bias towards specific types of events.[23][24] Trigger tool methodologies have shown promise for increasing AE detection, but still require targeted manual chart review, requiring increased resource utilization.[25][26] In our study, automated computer algorithms achieved the same or higher sensitivity and precision than the currently used trigger tool and voluntary incident-reporting approaches. In addition, the new algorithm was used to identify narcotic dosage, delivery and frequency

Table 4 Descriptive, IAA statistics (F-measure), and adverse event (AE)/medical error (ME) instance numbers of the annotated notes

	Procedure notes (pre-training)		Progress notes (training-1)		Progress notes (training-2)		Overall
Statistics							
Notes	1220		1453		590		3263
Patient days	1195		504		408		2107
Notes containing AE/ME	141 (11.6%)		77 (5.3%)		61 (10.3%)		279
Patients	395		16		17		428
	IAA (F)	AE/ME	IAA (F)	AE/ME	IAA (F)	AE/ME	AE/ME
Airway AEs							
Airway trauma	43.2%	39	**72.7%**	6	50.0%	4	49
ETT malposition	76.0%	36	60.9%	12	**100.0%**	3	51
Pneumothorax	18.0%	65	**71.7%**	16	0.0%	1	82
Unplanned extubation	70.4%	31	78.6%	44	**96.0%**	26	101
Vent assoc pneumonia	0.0%	2	**42.1%**	26	NULL	0	28
Other	18.2%	15	38.1%	2	**72.7%**	7	24
Medication/fluid AE/ME							
Wrong dose	N/A	0	N/A	0	N/A	0	0
Wrong medication	N/A	1	N/A	0	N/A	0	1
Medication reaction	N/A	6	N/A	6	N/A	0	12
Severe IV infiltrate	N/A	0	N/A	2	N/A	3	5
Code event	45.1%	32	**83.6%**	20	64.5%	15	67
All types	N/A	227	N/A	134	**N/A**	59	420

The same error events may be mentioned multiple times in the same notes or separate notes of the same day. The percentage is the percentage of identified AE/ME instances in the reviewed notes.
IAA, inter-annotator agreement.

discrepancies, a step that can be applied to detect dosage and frequency errors for other high-risk medications in the future. Neither the trigger tools nor the voluntary incident reports are well suited for this purpose.

The proposed algorithms have limitations. Extraction of all of the relevant data points from a complex, production EHR environment is a challenge and may be different in various EHR systems, which may limit application of the algorithms in other institutions.

The algorithms detect minor dosing discrepancies that differ from the exact ordered dose, but are not clinically relevant dosing errors. This becomes particularly important when thinking about the real-time use of error identification algorithms in the clinical arena. The algorithms also detect discrepancies between medication orders and deliveries when the medication is delivered based on a verbal order, and the computerized order follows. To address this common clinical scenario, we allowed a 15 min time discrepancy between medication delivery and medication order, but this time window may need to be adjusted for different types of medications and clinical situations.

Finally, the algorithms rely on documentation of medications and events within EHRs. Some errors may simply not be charted in the EHR, resulting in missing data. High-risk situations such as code events or anesthesia may also be recorded on paper, and associated AEs/MEs during such events may not be accurately reflected in EHRs. We did not review paper charting or anesthesia records in our current research. Similarly, errors due to pump programming are not detected in the EHR. For example, three narcotic delivery errors identified in the incident reports were due to pump programming errors. Currently, no data are recorded on the operation of the medication infusion pump because it is performed at the bedside, but future work should address infusion pump monitoring in error identification.

AE/ME annotation

We demonstrated improvement in IAA and the ability to recognize multiple error types through AE/ME annotation. There are limitations to using a NLP approach to error identification. High IAA requires precise definition of the annotation guidelines, which can be difficult to achieve for complex clinical situations. Multiple phrases may be used to identify the same event due to lack of standardization in the clinical notes, and inconsistencies exist amongst providers and over time. Several iterations were required to create precise definitions and identify the breadth of phrases representing one event, but unique descriptors may still be missed.

Clinical notes may also be biased in the types of AE/ME they reflect. For example, we found that medication/fluid AE/ME were rarely documented in clinical notes and were primarily detected in the structured data. As noted above, annotation of events also relies on the complete recording of events in the EHR, which may not comprehensively reflect all AE/MEs.

Finally, our current reported annotated results are AE/ME instances instead of unique events and can therefore be counted more than once. For example, if an unplanned extubation was mentioned more than once in a single note or the same AE/ME was recorded in multiple notes then we counted it more than once. Post-processing rules to convert annotated AE/ME instances to unique AE/ME events will need to be implemented in future work.

CONCLUSION

We developed a novel, EHR-based, phenotyping algorithm to detect NICU AE/ME. We evaluated the technical effectiveness of two AE/ME-specific algorithms on 1 year of retrospective EHR data and showed that they produced equivalent or higher sensitivity and precision than the current manual trigger tool and voluntary incident-reporting approaches. Using the algorithms

we were able to identify previously undetected or unreported AEs/MEs, demonstrating that the algorithms are a useful way to automate error detection in the neonatal intensive care environment.

REFERENCES

1 Solti I. EHR-Based patient safety: automated error detection in neonatal intensive care. 2012. http://projectreporter.nih.gov/project_info_details.cfm?aid=8334934&icde=15591108 (accessed 12 Mar 2013).

2 Das S, Eisenberg LD, House JW, et al. Meaningful use of electronic health records in otolaryngology: recommendations from the American Academy of Otolaryngology–Head and Neck Surgery Medical Informatics Committee. Otolaryngol Head Neck Surg 2011;144:135–41.

3 Blumenthal D. Launching HITECH. N Engl J Med 2010;365:382–5.

4 Classen DC, Resar R, Griffin F, et al. Global trigger tool' shows that adverse events in hospitals may be ten times greater than previously measured. Health Aff (Millwood) 2011;30:581–9.

5 Naessens JM, Campbell CR, Huddleston JM, et al. A comparison of hospital adverse events identified by three widely used detection methods. Int J Qual Health Care 2009;21:301–7.

6 Govindan M, Van Citters AD, Nelson EC, et al. Automated detection of harm in healthcare with information technology: a systematic review. Qual Saf Health Care 2010;19:e11.

7 Raju TN, Suresh G, Higgins RD. Patient safety in the context of neonatal intensive care: research and educational opportunities. Pediatr Res 2011;1:109–15.

8 Sharek PJ, Horbar JD, Mason W, et al. Adverse events in the neonatal intensive care unit: development, testing, and findings of an NICU-focused trigger tool to identify harm in North American NICUs. Pediatrics 2006;118:1332–40.

9 Kirkendall ES, Kloppenborg E, Papp J, et al. Measuring adverse events and levels of harm in pediatric inpatients with the Global Trigger Tool. Pediatrics 2012;130:e1206–14.

10 Snijders C, van Lingen RA, Molendijk A, et al. Incidents and errors in neonatal intensive care: a review of the literature. Arch Dis Child Fetal Neonatal Ed 2007;92:F391–8.

11 Zhan C, Miller MR. Administrative data based patient safety research: a critical review. Qual Saf Health Care 2003;12(Suppl 2):ii58–63.

12 Patient Safety Indicators (PSIs). http://www.qualityindicators.ahrq.gov/modules/psi_overview.aspx (accessed 18 Mar 2013).

13 McGlynn EA, Asch SM, Adams J, et al. The quality of health care delivered to adults in the United States. N Engl J Med 2003;348:2635–45.

14 Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc 2005;12:448–57.

15 Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. J Am Med Assoc 2011;306:848–55.

16 Sharek PJ, Classen D. The incidence of adverse events and medical error in pediatrics. Pediatr Clin North Am 2006;53:1067–77.

17 Bates DW, Boyle DL, Vander Vliet MB. Relationship between medication errors and adverse drug events. J Gen Intern Med 1995;10:199–205.

18 Nebeker JR, Barach P, Samore MH. Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Ann Intern Med 2004;140:795–801.

19 Snijders C, van Lingen RA, Klip H, et al. Specialty-based, voluntary incident reporting in neonatal intensive care: description of 4846 incident reports. Arch Dis Child Fetal Neonatal Ed 2009;94:210–15.

20 Muething SE, Conway PH, Kloppenborg E, et al. Identifying causes of adverse events detected by an automated trigger tool through in-depth analysis. Qual Saf Health Care 2010;19:435–9.

21 Risk MonitorPro®. http://www.rlsolutions.com/ (accessed 23 Mar 2013).

22 Ogren P. Knowtator: a protégé plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume: demonstrations; New York, New York: 2006, 273–5.

23 Noble DJ, Pronovost PJ. Underreporting of patient safety incidents reduces health care's ability to quantify and accurately measure harm reduction. J Patient Saf 2010;6:247–50.

24 Taylor JA, Brownstein D, Christakis DA, et al. Use of incident reports by physicians and nurses to document medical errors in pediatric patients. Pediatrics 2004;114:729–35.

25 Resar RK, Rozich JD, Simmonds T, et al. A trigger tool to identify adverse events in the intensive care unit. Jt Comm J Qual Patient Saf 2006;32:585–90.

26 Rozich JD, Haraden CR, Resar RK. Adverse drug event trigger tool: a practical methodology for measuring medication related harm. Qual Saf Health Care 2003;12:194–200.