Childhood Cancer Data Initiative Webinar Series

# Promises and Challenges: Automated Extraction of Electronic Health Record Data

*Tamara P. Miller, M.D., M.S.C.E. and Edward M. Krause, M.S.*

NIH NATIONAL CANCER INSTITUTE

May 21, 2024

# Today's Speakers

## Tamara P. Miller, M.D., M.S.C.E.

- Pediatric Oncologist, Aflac Cancer and Blood Disorders Center of Children's Healthcare of Atlanta
- Associate Professor of Pediatrics, Emory University School of Medicine

## Edward M. Krause, M.S.

- Data Integration Manager, Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia

# Agenda

1. Traditional methods of data collection for clinical research
2. Options for automated extraction of electronic health record (EHR) data for clinical research
3. Benefits and challenges of automated EHR data extraction
4. Use cases: Lessons learned from identifying, extracting, and using EHR data
5. Q&A

# Objectives

1. Review current options for extracting EHR data for research purposes
2. Recognize the benefits of using EHR data extracted in an automated way
3. Understand challenges with identifying, extracting, and using extracted EHR data and efforts to address these challenges

# Traditional Methods of Data Collection for Clinical Research

*Tamara P. Miller, M.D., M.S.C.E.*

# Data Collection for Clinical Research

- Data collected as part of clinical care can be leveraged for secondary research use
- Clinical data stored in EHRs (or previously, paper files)
  - Range of EHR vendors used across hospitals
  - A single institution may use multiple vendors for different data components
- Clinical research had typically relied on manually identifying and collecting (abstracting) data from EHRs
  - Clinical trials, development of cancer cohorts, retrospective cohort studies, multicenter data collection
- Significant challenges with manually abstracting data

# Challenges with Manual Data Abstraction

1. **Significant training is required** for abstractors to identify and collect data
   - Navigating EHR system
   - Detailed guidance on data ascertainment
     - Clinical knowledge of abstractors is variable and may be minimal
2. **Time consuming** to produce large volumes of data
3. Personnel required can be **expensive**
   - Requires clinician guidance, abstraction training, and personnel for data entry
4. **Challenging to coordinate** across multiple institutions
5. **Prone to human error**

# Manual Abstraction Can Lead to Inaccurate Ascertainment

- To better understand challenges with manual abstraction, we evaluated the accuracy of manual ascertainment of adverse events (AEs) on clinical trials as a use case
  - AE ascertainment performed by clinical research associates (CRAs) or research nurses (RNs)
  - Currently on trials, we manually identify and collect AEs from the EHR and report it into an electronic trial data capture system
- Quantified under-reporting using data from phase 3 cooperative oncology group clinical trial for de novo pediatric acute myeloid leukemia (AML)
  - Targeted 12 clinically important AEs
  - Compared the submitted AE reports on the trial to gold-standard physician chart abstraction at 14 hospitals across the U.S.

# Manual Abstraction Failed to Capture Data[1]

**Table 2.** Chart Abstraction Data Compared With Clinical Trial Adverse Event Report for Each of the 12 Grade 3 to 5 Toxicities

| Toxicity | Chart Abstraction, No. (%)* | Adverse Event Report | | | | |
|---|---|---|---|---|---|---|
| | | No. (%) | Sensitivity, % (95% CI) | Specificity, % (95% CI) | PPV, % (95% CI) | NPV, % (95% CI) |
| Hypertension | 28 (3.7) | 9 (1.2) | 21.4 (8.3 to 41.0) | 99.6 (98.8 to 99.9) | 66.7 (29.9 to 92.5) | 97.1 (95.6 to 98.2) |
| Hypotension | 46 (6.1) | 35 (4.6) | 56.5 (41.1 to 71.7) | 98.7 (97.6 to 99.4) | 74.3 (56.7 to 87.5) | 97.2 (95.8 to 98.3) |
| Hypoxia | 167 (22.0) | 30 (4.0) | 17.4 (12.0 to 24.0) | 99.8 (99.1 to 100) | 96.7 (82.8 to 99.9) | 81.0 (78.0 to 83.8) |
| ARDS | 13 (1.7) | 11 (1.5) | 38.5 (13.9 to 68.4) | 99.2 (98.3 to 99.7) | 45.5 (16.8 to 76.6) | 98.9 (97.9 to 99.5) |
| Anorexia | 307 (40.5) | 100 (13.2) | 30.6 (25.5 to 36.1) | 98.7 (97.1 to 99.5) | 94.0 (87.4 to 97.8) | 67.6 (63.9 to 71.2) |
| Typhlitis | 27 (3.6) | 11 (1.5) | 37.0 (19.4 to 57.6) | 99.9 (99.2 to 100) | 90.9 (58.7 to 99.8) | 97.7 (96.4 to 98.7) |
| DIC | 59 (7.8) | 7 (0.9) | 10.2 (3.8 to 20.8) | 99.9 (99.2 to 100) | 85.7 (42.1 to 99.6) | 92.9 (90.9 to 94.7) |
| VGS | 129 (17.0) | 103 (13.6) | 78.3 (70.2 to 85.1) | 99.7 (98.9 to 100) | 98.1 (93.2 to 99.8) | 95.7 (93.9 to 97.1) |
| IFI | 10 (1.3) | 10 (1.3) | 60.0 (26.2 to 87.8) | 99.5 (98.6 to 99.9) | 60.0 (26.2 to 87.8) | 99.5 (98.6 to 99.9) |
| Pain | 324 (42.7) | 56 (7.4) | 15.7 (12.0 to 20.2) | 98.9 (97.3 to 99.6) | 91.1 (80.4 to 97.0) | 61.1 (57.4 to 64.7) |
| Seizure | 5 (0.7) | 2 (0.3) | 0 (0.0 to 52.2) | 99.7 (99.0 to 100) | 0 (0.0 to 84.2) | 99.3 (98.5 to 99.8) |
| Renal failure | 6 (0.8) | 4 (0.5) | 50.0 (11.8 to 88.2) | 99.9 (99.3 to 100) | 75.0 (19.4 to 99.4) | 99.6 (98.9 to 99.9) |

**Sensitivity <50% for 8 of 12 targeted AEs**

**66% of AEs were missed**

**25% of submitted AEs were incorrect**

# Laboratory Adverse Events Are Also Inaccurately or Not Reported[2]


## Using electronic medical record data to report laboratory adverse events

Tamara P. Miller,[1,2,4] Yimei Li,[1,3] Kelly D. Getz,[1,2] Jesse Dudley,[4] Evanette Burrows,[4] Jeffrey Pennington,[4] Azada Ibrahimova,[5] Brian T. Fisher,[6,2,7,3] Rochelle Bagatell,[1,7] Alix E. Seif,[1,7] Robert Grundmeier[4,7] and Richard Aplenc[1,2,7,3]

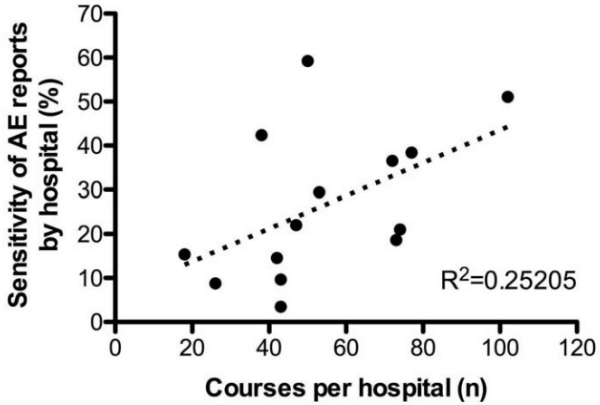
*[1]Division of Oncology, The Children's Hospital of Philadelphia, [2]Center for Pediatric Clinical Effectiveness, The Children's Hospital of*

**85% of lab AEs were missed**

**50% of submitted lab AEs were incorrect**

# Accuracy of AE Reporting on Clinical Trials[13]

- **Wide variability in sensitivity of AE reporting** that was not correlated with organ system
  - Differential false positive rates and missing data by AEs demonstrates widespread challenges with manual ascertainment of data
- **Variability between hospitals in accuracy of AE reporting**
  - Trend toward improved reporting at hospitals with more AML patients, but still missed/incorrect AEs
  - Raises concerns about comparability of data in multicenter studies



# Types of Manual Chart Abstraction Errors[2]

- Identified reasons for discrepancies in AE reports compared to gold-standard physician abstraction
- Manual transcription errors
  - Incorrect AE name selected (e.g., hypokalemia instead of hyperkalemia)
  - Laboratory data in EHR did not match AE reported
  - Apparent incorrect calculation compared to a reference range
- Reporting false positive results due to lack of clinical knowledge
  - E.g., Abnormal result normalized when re-checked within one hour
  - Selection of incorrect AE name when there are multiple similar options

# Variability in Abstraction Between Individuals[45]

- Even with guidance, individuals may take varying approaches
- Study mimicking trial AE reporting found that while some CRAs comprehensively report all AEs, others do not and only report the overarching syndromes
- Scharf and Colevas found lack of concordance between clinical trial data reported and Clinical Data Update System (CDUS) data submitted to FDA for the same patients
  - 27% of trial reported AEs not in CDUS
  - 28% of CDUS reported AEs not in clinical trial data

# Intrinsic Challenges of Manual Abstraction

- Inaccurate data collection may create false understanding of outcomes and patient experiences
  - Threatens trial integrity
- Reduces usefulness of clinical research for medical decision making
- Limits and creates biases in secondary analyses due to inaccuracies
  - Each additional analysis may require new data to be collected, which is time consuming and has same potential inaccuracies
- <u>New approach is needed:</u>
  - Automated extraction of data from the EHR has potential to address these challenges

# Options for Automated Extraction of EHR Data for Clinical Research

*Edward M. Krause, M.S.*

# Automated EHR Data Extraction

| Automated Data Extraction | Extract, Transform, and Load (ETL) |
| --- | --- |
| The use of a software tool to extract data from a system for downstream refinement or analysis (without manual oversight) | An automated software process where data is extracted from one or more sources, converted into a different format or structure, and loaded into a destination data store |

# Automated Extraction Method: Direct Database Access

- Relational database management system (RDBMS) comprised of data held in tables that are logically connected to each other through a data model
- Structured Query Language (SQL)
- Human-readable
- More granular/flexible access to data
- Greater flexibility to customize requested data elements
- Examples:
  - Oracle, SQL Server, PostgreSQL

## SQL Querv

```
SELECT
    p.pat_mrn_id "pat_id",
    ord.order_proc_id "order_id",
    to_char(res.result_time, 'YYYY-MM-DD') "result_date",
    to_char(res.result_time, 'HH24:MI') "result_time",
    eap.proc_name "procedure_name",
    co.name "component_name",
    res.ord_value "result",
    res.reference_unit "result_unit"
FROM
    patient p
    inner join order_proc ord
        on p.pat_id = ord.pat_id
    inner join clarity_eap eap
        on ord.proc_id = eap.proc_id
    left join order_results res
        on ord.order_proc_id = res.order_proc_id
    left join clarity_component co
        on res.component_id = co.component_id
WHERE
    (regexp_like(eap.proc_name,'^wbc|^white bl','i')
    or regexp_like(co.name,'^wbc|^white bl','i'))
    and ord.order_type_c = '7'
```

## Database Query Result

| pat_id | order_id | result_date | result_time | procedure_name | component_name | result | result_unit |
|---|---|---|---|---|---|---|---|
| 563453 | 68696078 | 2016-07-23 | 12:54 | CBC W/ DIFF | WBC | 4.7 | THOU/uL |
| 463452 | 35420070 | 2021-02-05 | 20:16 | CBC, PLATELET, W/ DIFF | NEUTROPHILS | 3247 | /uL |
| 574645 | 49844405 | 2023-08-22 | 5:45 | CBC, PLATELET, W/ DIFF | LYMPHOCYTES | 1.5 | x10E3/uL |
| 645634 | 78648696 | 2019-10-15 | 10:31 | CMPLT BLD CT WITH DIFF | PLATELET COUNT | 480 | THOU/uL |

# Automated Extraction Method: Application Programming Interface (API)

- A set of protocols or logical rules that allow software systems to communicate and exchange data in a structured manner
- An abstraction layer over a database machine readable
- Less granular/flexible access to data
- Available data elements pre-defined by system
- Examples:
  - API types: REST, SOAP
  - Data formats: JSON, XML, RDF

## API Request Response


```
{
  "fullUrl": "https://example.com/base/Observation/r7",
  "resource": {
    "resourceType": "Observation",
    "id": "r7",
    "status": "final",
    "code": {
      "coding": [
        {
          "system": "http://loinc.org",
          "code": "6690-2",
          "display": "Leukocytes [#/volume] in Blood by Automated count"
        }
      ],
      "text": "White Cell Count"
    },
    "subject": {
      "reference": "Patient/pat2"
    },
    "valueQuantity": {
      "value": 4.6,
      "unit": "x10*9/L",
      "system": "http://unitsofmeasure.org",
      "code": "10*9/L"
    },
    "referenceRange": [
      {
        "low": {
          "value": 4,
          "unit": "x10*9/L",
          "system": "http://unitsofmeasure.org",
          "code": "10*9/L"
        },
        "high": {
          "value": 11,
          "unit": "x10*9/L",
          "system": "http://unitsofmeasure.org",
          "code": "10*9/L"
        }
      }
    ]
  }
}
```

# Automated EHR Data Extraction Tools and Standards

- **Databases:**
  - Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) ETL Toolkit
  - ExtractEHR
- **APIs:**
  - Fast Healthcare Interoperability Resources (FHIR) API
  - Research Electronic Data Capture (REDCap) Clinical Data Interoperability Services

# Benefits and Challenges of Automated EHR Data Extraction

*Edward M. Krause, M.S.*

# Automated Data Extraction: Benefits

- **Standardization**
  - Extraction of parallel data elements for every patient
  - Does not rely on individual clinical knowledge of abstractor
  - Data harmonization as part of extraction coding
  - Reduced human error of incorrect selection of elements
- **Centrally managed software that can be created and deployed**
  - Upfront effort to identify elements and implement
  - Updates and dissemination controlled by central team members
- **Reusable processes**
  - Once installed, it can be reused repeatedly
- **Scalable processes**
  - Large volumes of data can be extracted more efficiently than at an individual level
  - Multiple institutions

# Automated Data Extraction: Challenges

- **Access to EHR systems/interfaces**
  - Security and privacy requirements; risk assessments
- **Knowledge of EHR data model**
  - Proprietary information
  - Complex entity relationships (18,000+ tables)
- **Multiple EHR products on the market**
- **Varying EHR data storage formats**
  - Structured, unstructured, and images/PDFs
- **Technical requirements for software implementation and use**
  - Programming knowledge
  - Configuration and deployment
  - Maintenance

# Software Tool: ExtractEHR

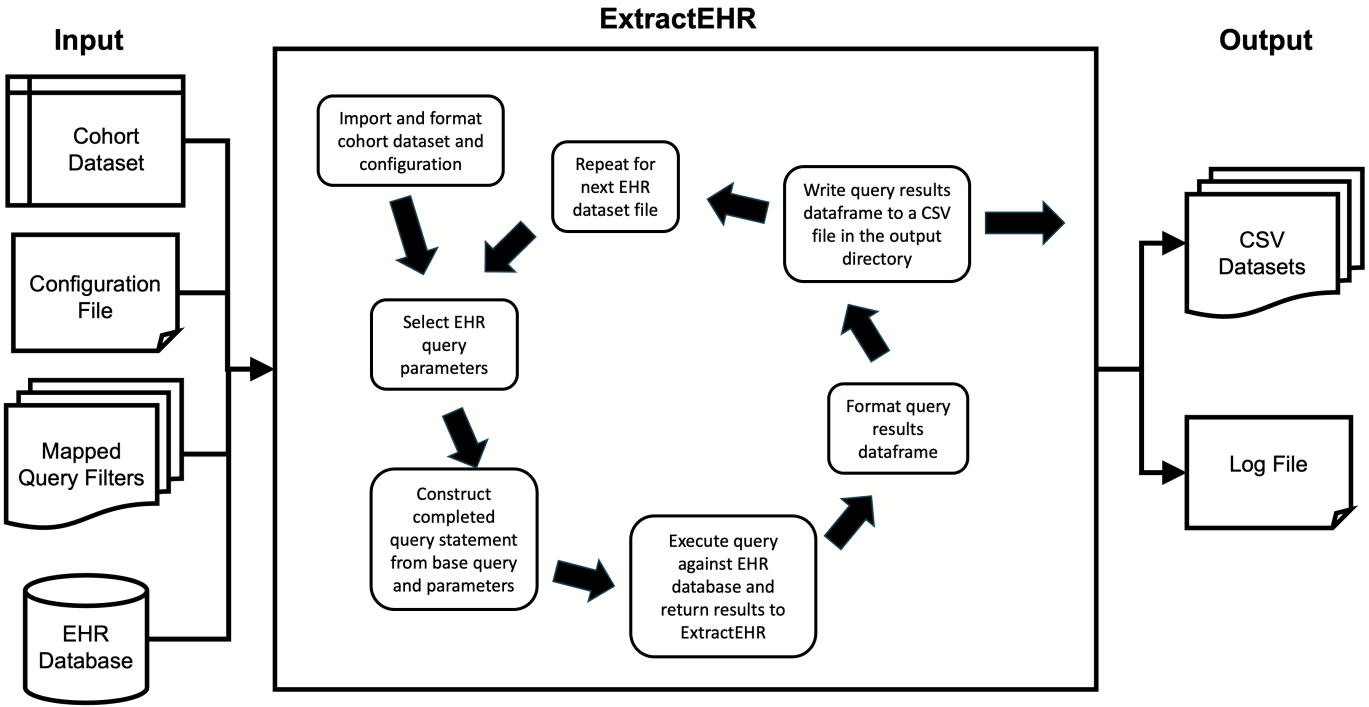
## Description:

- Written in the R programming language
- Implemented in the R Studio integrated development environment
- Utilizes direct database access and pre-defined SQL queries
- Code hosted in private GitHub repository
- Participating institutions supply:
  - EHR database credentials/connection
  - Patient cohort data set
  - Institution- and EHR-specific inclusion/exclusion criteria (query filters)

## Requirements:

- Staffing at each participating site
  - Investigator (clinical/research subject matter expert)
  - Data analyst/programmer (SQL, R, EHR data model)
  - Optional: Clinical research staff (coordinator, project manager, etc.)
- Software and data
  - Workstation/laptop with R and R Studio installed
  - ExtractEHR GitHub repository access
  - EHR database access (e.g., Epic Clarity)
  - Patient cohort data set

# ExtractEHR Diagram



# ExtractEHR-Mapped Data Sets

- Address history
- Demographics
- Encounters/visits
- Medication orders/administrations
- Procedures
- Vital signs
- Pathology reports
- Radiology reports
- Oncology clinical notes
- Genomic results
- Laboratory results (all labs or specific labs listed to the right)
  - Absolute blasts
  - Albumin
  - Alkaline phosphatase (ALP)
  - Alanine aminotransferase (ALT)
  - Amylase
  - Aspartate aminotransferase (AST)
  - Bicarbonate
  - Bilirubin
  - Blood urea nitrogen
  - Calcium
  - CD19 cell count
  - CD4 cell count
  - Chloride
  - Creatinine
  - Creatinine phosphokinase
  - C-reactive protein
  - Fibrinogen
  - Free T4
  - Gamma glutamyl transferase (GGT)
  - Glucose
  - Hematocrit
  - Hemoglobin
  - HDL cholesterol
  - Immunoglobulin-G
  - International normalized ratio (INR)
  - Lipase
  - LDL cholesterol
  - Lymphocytes
  - Magnesium
  - Methotrexate
  - Microbiology
  - Neutrophils
  - Partial thromboplastin time
  - Phosphorus
  - Platelets
  - Potassium
  - Pregnancy
  - Prothrombin time
  - Sodium
  - Thyroid stimulating hormone (TSH)
  - Total cholesterol
  - Total protein
  - Triglyceride
  - Troponin
  - Uric acid
  - Urinalysis
  - Urine creatinine
  - Urine protein
  - White blood cells

# Post-Extraction Data Processing

- EHR data can be extracted in one large group or limited to specific variables
- EHR data are collected for clinical purposes
  - Documentation useful for clinical purposes may not be sufficient or understandable for research
  - Data storage structures designed for system operations often not optimized for research
  - May need processing or transformation for research use

| name | taken_date | taken_time | dose | unit | infusion_rate | infusion_rate_unit | mar_action |
|---|---|---|---|---|---|---|---|
| METHOTREXATE IV SOLUTION | 5/23/19 | 11:32:16 | NA | NA | NA | NA | MAR Hold |
| METHOTREXATE IV SOLUTION | 5/23/19 | 13:02:48 | NA | NA | NA | NA | MAR Unhold |
| METHOTREXATE IV SOLUTION | 5/23/19 | 22:35:00 | 2.7 | g | 115 | mL/hr | New Bag |
| METHOTREXATE IV SOLUTION | 5/23/19 | 23:30:00 | 2.7 | g | 144 | mL/hr | Rate Change |
| METHOTREXATE IV SOLUTION | 5/24/19 | 0:16:00 | 2.7 | g | 115 | mL/hr | Rate Change |
| METHOTREXATE IV SOLUTION | 5/24/19 | 1:35:00 | 0 | g | 0 | mL/hr | Stopped |

# Post-Extraction Processing Tools: CleanEHR and GradeEHR

- **CleanEHR**
  - Software tool designed to process ExtractEHR-generated data sets
  - Removes false or duplicate results; standardizes data element formatting; creates cleaning summary metrics of each data set
- **GradeEHR**
  - Software tool designed to process CleanEHR-generated data sets to assign AE grades to laboratory results
  - Written to match Common Terminology Criteria for Adverse Events (CTCAE) definitions

# Use Cases

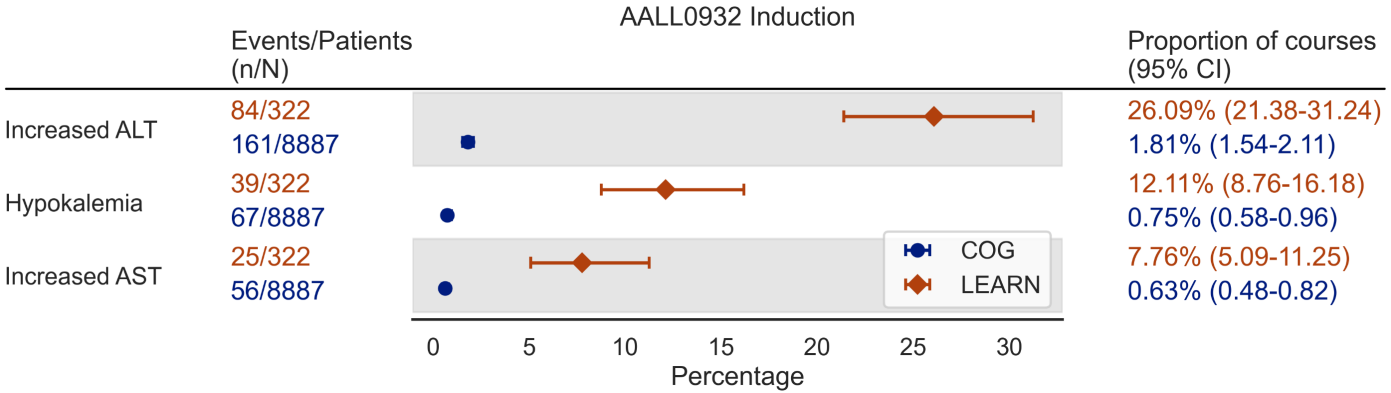
*Tamara P. Miller, M.D., M.S.C.E.*

# Use Case: Development of Cancer Cohorts

- EHR data can populate cancer cohorts to provide data to answer clinical epidemiology questions
- Implemented ExtractEHR to create a multicenter cohort for clinical research: Leukemia Electronic Abstraction of Records Network (LEARN)
  - Deployed at 4 institutions, 3 in process (Epic and Cerner EHR vendors)
  - Pediatric patients with de novo acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)
    - Trained manual abstraction collects clinical trial enrollment status and chemotherapy course dates, then run ExtractEHR
  - De-identified, extracted data shared and stored centrally
  - CleanEHR and GradeEHR applied to create processed data sets that can be used to answer research questions

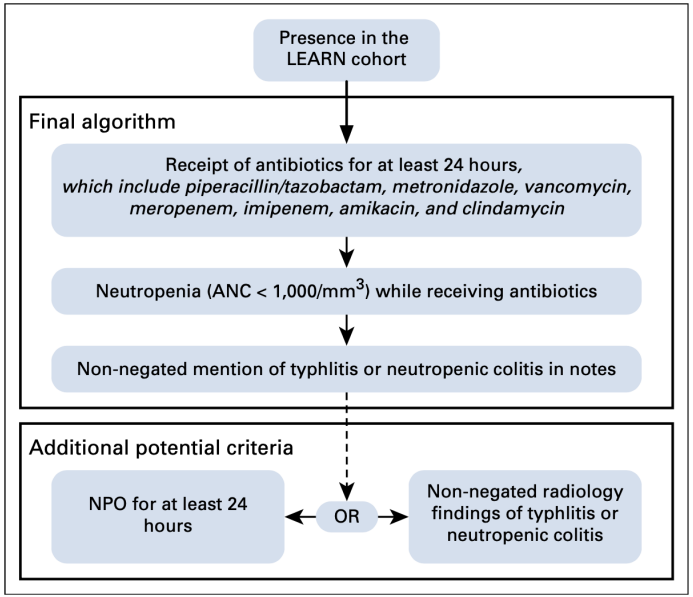
# Using EHR Data to Answer Clinical Epidemiology Questions[6]

- Described rates of laboratory AEs for children with ALL and AML
  - EHR data extracted for targeted labs, processed using CleanEHR and GradeEHR
  - Efficient, granular, accurate assessment across multiple hospitals
- More comprehensive than manual ascertainment on trials



# Use of EHR Data to Understand Complex Phenotypes[7]

- Can use extracted and processed EHR data to describe complex outcomes and phenotypes
- Example: AE of typhlitis
  - Extracted targeted data elements from EHR
  - Included structured and unstructured data that need processing and integration
  - Natural language processing (NLP) and machine learning applied to identify relevant AE data from free text fields
- After iterative development, algorithm successfully identified chemotherapy courses where a patient may have had typhlitis
  - Reduced number of courses needing review to confirm typhlitis occurred by 96% (961→37)



# Use Case: Registries

- Extracting raw data for transfer to Surveillance, Epidemiology, and End Results (SEER) registry
- EHR variables extracted in one large group without post-extraction processing
  - Hospital encounters, laboratory test results, medications, procedures, vital signs, radiology reports, pathology reports, oncology clinical notes
- Successful at Children's Healthcare of Atlanta with transfer to Georgia Cancer Registry: Initial extraction of 306 patients; 3,015,822 data elements
  - Three additional sites in process
- Provides granular data formatted in parallel across hospitals for SEER use
  - Post-extraction processing can be performed as needed for registry projects

# Use Case: Childhood Cancer Data Initiative (CCDI)

- **Goal:** Supplement molecular data with EHR data to provide clinical context
- **Current pilot:** Use ExtractEHR to provide treatment and outcome data for Children's Brain Tumor Network (CBTN) patients with molecular data in CCDI
  - Extracting data from two large hospitals with post-extraction processing and cleaning
    - Reusing ExtractEHR code from LEARN implementation
    - Performing significant post-extraction processing (CleanEHR) to create clean data set describing clinical experience during cancer treatment
  - Example: Identification of all chemotherapy agents, doses, regimens received
  - Two approaches: Standard ExtractEHR package and FHIR-based version
  - Data will:
    - Benefit CBTN by reducing current manual collection of data
    - Expand the level of detail of clinical data in the CCDI Data Ecosystem

# Use Case: Clinical Trials

- Pilot study PEPN21EHR/PBTCN-15 (NCT05020951) within the Children's Oncology Group (COG) Pediatric Early Phase Clinical Trials Network and Pediatric Brain Tumor Consortium (PBTC)
  - **Goal:** Automatically extract EHR data and directly import into trial electronic data capture system (Medidata Rave) across institutions
  - Data extracted via ExtractEHR or locally-developed extraction methods
    - Raw data included with post-extraction formatting to match Rave requirements
  - Successful extraction, formatting, and upload at all seven participating sites
- Permits comprehensive and efficient data capture for trial-based research
- Lessons learned about operational needs to permit an automated process

# Summary

- Intrinsic challenges in manual data collection can be addressed by automated EHR data capture
- EHR data extraction is feasible, but requires upfront effort to implement
  - Once implemented, extraction tools are repeatable and reusable for other use cases
  - Extracted data can be formatted using code to match use case needs
- EHR data need post-extraction processing for use in clinical research
- Once extraction and processing pipelines are created, a wide range of clinical research use cases are possible

# References


- [1]Miller TP, Li Y, Kavcic M, et al. Accuracy of adverse event ascertainment in clinical trials for pediatric acute myeloid leukemia. *Journal of Clinical Oncology* 2016 Feb 16. doi: 10.1200/JCO.2015.65.5860.
- [2]Miller TP, Li Y, Getz KD, et al. Using electronic medical record data to report laboratory adverse events. *British Journal of Haematology* 2017 Feb 1. doi: 10.1111/bjh.14538.
- [3]Miller TP, Li Y, Kavcic M, et al. Center-level variation in accuracy of adverse event reporting in a clinical trial for pediatric acute myeloid leukemia: a report from the Children's Oncology Group. *Haematologica* 2017 Sep. doi: 10.3324/haematol.2017.168815.
- [4]Miller TP, Fisher BT, Getz KD, et al. Unintended consequences of evolution of the Common Terminology Criteria for Adverse Events. *Pediatric Blood & Cancer* 2019 Apr 9. doi: 10.1002/pbc.27747.
- [5]Scharf O, Colevas AD. Adverse event reporting in publications compared with sponsor database for cancer clinical trials. *Journal of Clinical Oncology* 2006 Aug 20. doi: 10.1200/JCO.2005.05.3959.
- [6]Miller TP, Getz KD, Li Y, et al. Rates of laboratory adverse events by course in paediatric leukaemia ascertained with automated electronic health record extraction: a retrospective cohort study from the Children's Oncology Group. *The Lancet Haematology* 2022 Jul 20. doi: 10.1016/S2352-3026(22)00168-5.
- [7]Miller TP, Li Y, Masino AJ, et al. Automated ascertainment of typhlitis from the electronic health record. *JCO Clinical Cancer Informatics* 2022 Oct 5. doi: 10.1200/CCI.22.00081.

# Q&A

# How You Can Engage with CCDI

**Learn about CCDI, register for upcoming events, and subscribe to our monthly newsletter:**
cancer.gov/CCDI

**Access CCDI data and resources:**
ccdi.cancer.gov

**Questions? Email us at:**
NCIChildhoodCancerDataInitiative@mail.nih.gov

# Thank you for attending!


cancer.gov

cancer.gov/espanol