High-throughput Mining of Electronic Medical Records Using Generalizable Autonomous Scripts

Ryan H. Rochat MD, PhD, GEMS[1], Gail J. Demmler-Harrison MD[1]

(Poster #132)

Affiliations: 1. Baylor College of Medicine , Texas Children's Hospital, Department of Pediatrics, Houston, Texas, USA Mailing Address: 1120 Bates Ave, Houston, Texas, 77030 Email Address: rochat@bcm.edu

Background:

The Electronic Medical Record (EMR) has become a modern compendium of health information, from broad clinical assessments down to an individual's heart rate. The wealth of information in these EMRs hold promise for clinical discovery and hypothesis generation. Unfortunately, as these systems have become more robust, mining them for relevant clinical information is hindered by the overall data architecture, and often requires the expertise of a clinical informatician to extract relevant data. However, as the information presented to the clinician through the digital workspace is derived from the core EMR database, the format is well structured and can be mined using text recognition and parsing scripts.

As clinicians learn to work within the confines of their own EMR, they have an adept sense of where the salient data is stored within these systems and how to access the data through their own specific Graphic User Interfaces (GUIs). Accordingly, individual users are often the best at identifying the data that is most relevant to their area of expertise (e.g. antimicrobial susceptibility for the infectious disease physician, ventilator settings for the critical care physician). By capitalizing on this knowledge, it is possible to write a programs which can be generalized to extract this data on a case-by-case basis for the provider.

Methods:

Here we present a program which can parse output from Epic® Hyperspace, generating a relational database of clinical information. To facilitate ease of use, our protocol capitalizes on the familiarity of Microsoft Excel® as an intermediary for storing the raw output from the EMR, with data parsing and processing scripts written in SAS V9.4 (Cary, N.C.). Patients that fit specific user defined clinical criteria can be identified using most EMR auditing tools. We used SlicerDicer™, an auditing tool included with our release of Epic®, to identify a unique subset of patients to extract medical records from. The "SnapShot" view of these patients contains user-selectable text that when copied retains the inherent data structure of the viewing window. This approach can be used to manually extract large amounts of data from the EMR in a rich-text-format. The resulting text is structured in the sense that all relevant metadata, which describse the free text, can be used to direct naming of variable and assigning values. For each data-stream, examples of the data parsing is shown in the accompanying figures.

Generalizability:

One focus of this project was to develop a program that could adapt to its input and produce a relational database that contains all of the salient features from a relatively diverse data-stream. For example, if only chemistry laboratory values were extracted from the EMR, this process would recognize and collect only those variable that were extracted, associating them with the patient's from whose chart they were obtained. Furthermore, by only requiring the user to select those values they desire the resulting database would be tailored to that specific clinical question. By capitalizing on headers present in the data-stream, it is possible to identify data relating to variable names, as opposed to specific values.

Definitions:

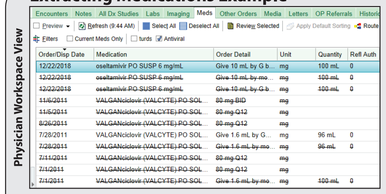
EMR (Electronic Medical Record): A systematized collection of personal health information, digitally stored.
GUI (Graphical User Interface): An interface designed for a user to interact with an electronic device through a series icons and text based commands.
Delimiter: A sequence of characters used to separate specific elements of data.
Concatenation: The process of joining strings of characters together to form a longer string of data.
Parsing: The process of deconvoluting strings of data, thus extracting information separated by elimiters.

Theory:

Information entered into the EMR is stored in a centralized database where the individual pieces of information can be related back to each other through a series of unique identifiers. When accessing the medical record, the GUI pulls specific elements from this database and displays them in a format recognizable to the operator. The information that is displayed to the user is visually partitioned, and this is manifest typographically in the raw data as a series of unique characters (i.e. delimiters). Knowledge of which delimiters separate specific data, makes it possible to extract this information by parsing long streams of text for these characters.

Extracting Medications Example

Physician Workspace View



Raw Data Extraction

```
Rx #: AW011115
VALGANciclovir (VALCYTE) PO SOLN 50 mg/mL [54463830] DISCONTINUED
Order Details
Dose: 80 mg Route: Per G button Frequency: EVERY 12 HOURS
Dispense Quantity: 96 mL Refills: 0 Fills Remaining: 1
Sig:
Give 1.6 mL by mouth every 12 hours for 30 days.
Discontinue Date: 12/15/2011 1558 Discontinue User: XXX, XXX Discontinue Reason: --
Written Date: 07/28/11 Expiration Date: 07/27/12
Start Date: 07/28/11 End Date: 12/15/11
```

SAS V9.4 Code

```
data temp2;
set temp;
format start_date end_date mmddyy10.;
text=compbl(A||B||C||D||E||F||G||H||I);
if index(text,'(CDM: N/A)') then mrn=input(scan(scan(text,2,'#'),1,' '),best12.);
if index(text,'Order ')>0 then do;
medication=scan(text,1,' ');
order_number=input(compress(scan(text,1,'(')*'b'),,'dk'),best12.);
end;
if index(text,'Start Date:') then start_date=input(substr(text,index(text,'Start
Date:')+12,8),mmddyy8.);
if index(text,'End Date:') then end_date=input(substr(text,index(text,'End
Date:')+10,8),mmddyy8.);
if index(text,'Start:') then start_date=input(substr(text,index(text,'Start:')+7,8),mmddyy8.);
if index(text,'End:') then end_date=input(substr(text,index(text,'End:')+5,8),mmddyy8.);
if index(A,'Dose:')>1 then dose=scan(scan(text,2,':'),1,' ')||" "||scan(scan(text,2,':'),2,
if index(B,'Route:')>1 then route=scan(B,2,':');
if index(C,'Frequency:')>1 then frequency=scan(C,2,':');
if index(text,'Routing History Report')>1 then output=1;
else output=0;
if output=1;
keep medication order_number start_date end_date dose route frequency mrn;
retain mrn medication order_number start_date end_date dose route frequency;
run;
```

Data Concatenation

Line 1: Rx #: AW011115
Line 2: Rx #: AW011115@VALGANciclovir (VALCYTE) PO SOLN 50 mg/mL [54463830] DISCONTINUED
Line 3: Rx #: AW011115@VALGANciclovir (VALCYTE) PO SOLN 50 mg/mL [54463830] DISCONTINUED@Order Details
....
Line 10: Rx #: AW011115@VALGANciclovir (VALCYTE) PO SOLN 50 mg/mL [54463830] DISCONTINUED@Order Details@Dose: 80 mg Route: Per G button Frequency: EVERY 12 HOURS Dispense Quantity: 96 mL Refills: 0 Fills Remaining: 1@Sig:@Give 1.6 mL by mouth every 12 hours for 30 days.@Discontinue Date: 12/15/2011 1558 Discontinue User: XXX, XXX A Discontinue Reason: -- @Written Date: 07/28/11 Expiration Date: 07/27/12@Start Date: 07/28/11 End Date: 12/15/11

Data Parsing

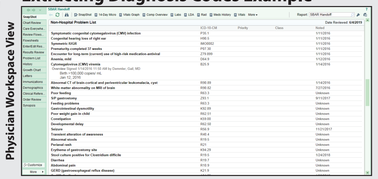
Variable	Input	Delimiter (position)	Output
order_number	Line 10	"@" (2) **then** "[" (1)	54463830
medication	Line 10	"@" (2) **then** "(" (1)	VALGANciclovir
dose	Line 10	"@" (4) **then** ":" (2) **then** " " (1 & 2)	80 mg
start_date	Line 10	"@" (10) **then** ":" (2) **then** " "(1)	07/28/11
end_date	Line 10	"@" (10) **then** ":" (3)	12/15/11

Database Output

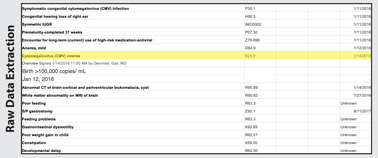
Obs	medication	dose	route	frequency	start_date	end_date	order_number
20	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	08/26/2011	08/31/2011	55421469
21	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	11/05/2011	11/06/2011	57992042
22	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	11/06/2011	11/07/2011	57996675
23	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	06/03/2011	07/01/2011	52804036
24	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	06/04/2011	06/03/2011	52804035
25	VALGANciclovir	80 mg	Oral	EVERY 12 HOURS	06/03/2011	06/03/2011	52804031
26	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/06/2011	05/10/2011	51852204
27	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/06/2011	05/06/2011	51852203
28	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/11/2011	05/16/2011	52026531
29	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/17/2011	05/23/2011	52145999
30	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/24/2011	05/27/2011	52344455
31	ganciclovir	80 mg	Oral	EVERY 12 HOURS	05/28/2011	06/03/2011	52642594

Extracting Diagnosis Codes Example

Physician Workspace View



Raw Data Extraction



SAS V9.4 Code

```
/*clean up the data from xls file that was copy/pasted from EPIC Overview page*/
data temp;
format mrn num best12. retain_diagnosis $500. comments $500. date_fix mmddyy10.;
set raw_diagnosis;
/*MRN occurs only once for a list of problems, so retain it for all problems*/
mrn_num=input(mrn,best12.);
if mrn_num^=. then retain_mrn=mrn_num;
if mrn_num= then mrn_num=retain_mrn;
/*fix any training spaces in the icd10 code, also carry the icd10 code to subsequent*/
/*lines when reading in the data to parse additional diagnoses' into a comments section.*/
/*have to do this because only have one icd10 code and then additional entries for comments*/
/*associated with the diagnosis.*/
if compress(icd10)^='' then do;
retain diagnosis='';
epic_diagnosis=diagnosis;
end;
else do;
retain diagnosis=compbl(retain_diagnosis)||diagnosis;
epic_diagnosis=epic_diagnosis;
end;
/*essentially do the same thing as above to make sure that the entered date is in*/
/*the record completely. FYI date added is probably bad for most kids >8yrs old*/
/*at the DOB did not exist then*/
```

Data Concatenation

Line 1: Cytomegalovirus (CMV) viremia#B25.9#1/14/2016
Line 2: Cytomegalovirus (CMV) viremia#B25.9#1/14/2016@Overview Signed 1/14/2016 11:50 AM by Demmler, Gail, MD
Line 3: Cytomegalovirus (CMV) viremia#B25.9#1/14/2016@Overview Signed 1/14/2016 11:50 AM by Demmler, Gail, MD@Birth >100,000 copies/ mL
Line 4: Cytomegalovirus (CMV) viremia#B25.9#1/14/2016@Overview Signed 1/14/2016 11:50 AM by Demmler, Gail, MD@Birth >100,000 copies/ mL@Jan 12, 2016
...
Line 8: Cytomegalovirus (CMV) viremia#B25.9#1/14/2016@Overview Signed 1/14/2016 11:50 AM by Demmler, Gail, MD@Birth >100,000 copies/ mL@Jan 12, 2016 @Abnormal CT of brain-cortical and periventricular leukomalacia, cyst#R90.89#1/14/2016@White matter abnormality on MRI of brain#R90.82#7/27/2016@Poor feeding#R63.3#Unknown@S/P gastrostomy#Z93.1#9/11/2017

Data Parsing

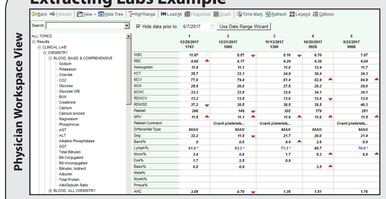
Variable	Input	Delimiter (position)	Output
epic_diagnosis	Line 8	"@" (1) **then** "#" (2)	Cytomegalovirus (CMV) viremia
icd10	Line 8	"@" (1) **then** "#" (2)	B25.9
date	Line 8	"@" (2) **then** "#" (3)	1/14/2016
epic_diagnosis	Line 8	"@" (8) **then** "#" (1)	S/P gastrostomy

Database Output

icd10	epic_diagnosis	diagnosis_age
Enterocolitis due to Clostridium difficile, not specified as recurrent	C. difficile diarrhea	2.12457
Cytomegaloviral disease, unspecified	Cytomegalovirus (CMV) viremia	0.07940
Other specified bacterial agents as the cause of diseases classified elsewhere	Clostridium difficile infection	2.13005
Anemia, unspecified	Anemia, mild	0.07392
Other neutropenia	Autoimmune neutropenia	2.47502
Neutropenia, unspecified	Febrile neutropenia	2.13552
Developmental disorder of speech and language, unspecified	Speech/language delay	2.13279
Sensorineural hearing loss, unilateral, right ear, with unrestricted hearing on the contrala	Sensorineural hearing loss (SNHL) of right ear with unrestricted hearing of left ear	2.13279
Unspecified sensorineural hearing loss	Congenital hearing loss of right ear	0.07118
Unknown	Symmetric IUGR	0.07118
Gastro-esophageal reflux disease without esophagitis	GERD (gastroesophageal reflux disease)	
Constipation, unspecified	Constipation	
Other specified diseases of the digestive system	Gastrointestinal dysmotility	
Other complications of gastrostomy	Erythema of gastrostomy site	
Preterm newborn, unspecified weeks of gestation	Prematurity-completed 37 weeks	0.07118

Extracting Labs Example

Physician Workspace View



Raw Data Extraction

```
6/7/17 17:56
WBC: 10.38
Hemoglobin: 15.7
Platelet: 159
Seg: 28.4 (L)
Band%: 20.7 (H)
Lymph%: 37.9 (H)
Mono%: 6.9
Eos%: 1.7
Baso%: 0.9
ANC: 5.10 (L)

6/9/17 3:45
WBC: 9.54
Hemoglobin: 13.9 (L)
Platelet: 153

7/28/17 2:02
WBC: 7.80
Hemoglobin: 10.3 (L)
Platelet: 327
Seg: 22.1
Band%: 1.6
Lymph%: 73.8
Mono%: 2.5
ANC: 1.85

9/7/17 9:32
WBC: 6.71
Hemoglobin: 11.5
Platelet: 256
Seg: 19.8 (L)
Band%: 0.8

11/13/17 13:00
WBC: 5.16 (L)
Hemoglobin: 11.8
Platelet: 202
Seg: 21.7
Band%: 4.4 (H)
Lymph%: 71.3
Mono%: 1.7
Eos%: 0.9
ANC: 1.35

12/21/17 10:05
WBC: 5.57 (L)
Hemoglobin: 11.1
Platelet: 149 (L)
Seg: 11.8 (L)
```

SAS V9.4 Code

```
data temp_hematology;
set all_data;
format variable $20. date mmddyy10. time time5.;
if length(compress(A))=1 or index(A,'(MRN ')>0 then count=1;
if index(A,'Results for')=1 then do;
MRN=input(substr(A,index(A,'(MRN ')+4,index(A,')')-index(A,'(MRN ')-4),best12.);
end;
if count=2 then do;
date_all=put((A-21916)*86400,datetime.);
time=input(substr(date_all,9,8),time10.);
date=input(substr(date_all,1,7),date.);
end;
if count>2 then do;
variable=scan(A,1,':');
if index(variable,'%') then variable=substr(variable,1,index(variable,'%')-1)||'_pct';
value=input(compress(scan(A,2,':'),'(').'a'),best12.);
end;
count=count+1;
if variable='' and value'=.;
retain MRN count date all date time;
run;
```

Data Concatenation

Line 1: 6/7/17 17:56
Line 2: 6/7/17 17:56@WBC: 10.38
Line 3: 6/7/17 17:56@WBC: 10.38@Hemoglobin: 15.7
Line 4: 6/7/17 17:56@WBC: 10.38@Hemoglobin: 15.7@Platelet: 159
...
Line 16: 6/7/17 17:56@WBC: 10.38@Hemoglobin: 15.7@Platelet: 159@Seg: 28.4 (L)@Band%: 20.7 (H)@Lymph%: 37.9 (H)@Mono%: 6.9@Eos%: 1.7@Baso%: 0.9@ANC: 5.10 (L)@6/9/17 3:45
Line 17: 6/7/17 17:56@WBC: 10.38@Hemoglobin: 15.7@Platelet: 159@Seg: 28.4 (L)@Band%: 20.7 (H)@Lymph%: 37.9 (H)@Mono%: 6.9@Eos%: 1.7@Baso%: 0.9@ANC: 5.10 (L)@6/9/17 3:45@WBC: 9.54@Hemoglobin: 13.9 (L)@Platelet: 153
Line 18: 6/7/17 17:56@WBC: 10.38@Hemoglobin: 15.7@Platelet: 159@Seg: 28.4 (L)@Band%: 20.7 (H)@Lymph%: 37.9 (H)@Mono%: 6.9@Eos%: 1.7@Baso%: 0.9@ANC: 5.10 (L)@6/9/17 3:45@WBC: 9.54@Hemoglobin: 13.9 (L)@Platelet: 153@Seg: 54.6

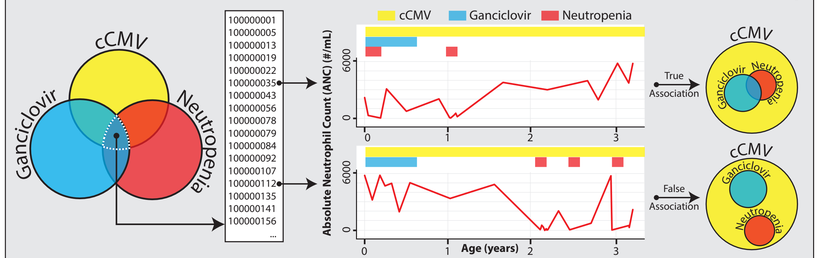
Data Parsing

Variable	Input	Delimiter (position)	Output
date	Line 18	"@" (1) **then** " " (1)	6/7/17
time	Line 18	"@" (1) **then** " " (2)	17:56
lab	Line 18	"@" (2) **then** ":" (1)	WBC
lab_value	Line 18	"@" (2) **then** ":" (2)	10.38

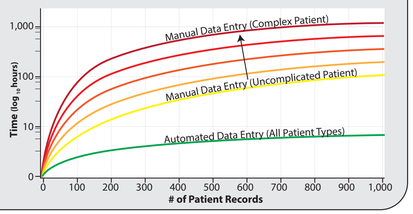
Database Output

date	time	age_value	WBC	Hemoglobin	neutrophils_pct	Lymph_pct	Band_pct	Baso_pct	Mono_pct	ANC
01/11/2016	11:59	0.07118	10500	9.7	27.8	65.9	.	0.4	4.2	5838.0
01/28/2016	11:41	0.11773	13300	10.3	16.8	77.2	.	0.2	5.1	4468.8
02/11/2016	11:31	0.15606	11300	11.2	14.3	79.8	.	0.9	4.2	3231.8
03/15/2016	11:16	0.24641	9600	10.9	30.5	64.5	.	0.6	3.8	5856.0
04/06/2016	11:40	0.30664	9200	10.9	25.7	70.8	.	0.5	2.1	4728.8
05/04/2016	10:53	0.38330	10000	11.1	25.0	72.0	.	0.0	2.0	5000.0
06/07/2016	10:27	0.47639	10400	11.4	19.0	77.9	.	0.2	2.0	1976.0
07/22/2016	11:49	0.59959	12600	10.6	20.1	73.2	.	0.2	5.6	5065.2
01/11/2017	11:07	1.07324	10600	9.3	16.0	71.4	.	0.6	11.4	3392.0
07/19/2017	12:17	1.50069	10900	11.3	44.7	49.2	.	0.3	4.9	4880.0
01/29/2018	14:50	2.12183	6500	12.1	2.0	84.3	.	0.3	11.9	130.0
01/31/2018	10:30	2.12731	4640	12.3	.	86.6	1.7	.	10.0	80.0

Audit Level Query (All Patients) / Medical Record Numbers / Detailed Medical Records (Individual Patients) / Diagnostic Relationship



Objective	Technique: Manual Entry	Technique: Our Approach
Extract ICD-10 Diagnosis Codes	25 hours	20 minutes
Extract Laboratory Data	90 hours	30 minutes
Extract Medication Administration Data	25 hours	40 minutes
Total Duration	**140 hours**	**1.5 hours**



Results:

We extracted the diagnosis codes, standard laboratory values, and antiviral medications for 190 patients seen in our Congenital Cytomegalovirus Clinic. Manual extraction of this data into Microsoft Excel® took 1 hour, and the scripts to parse the data took less than 5 seconds to run. Data from these patients included: 3800 lines of ICD-10 codes (along with their metadata), 33,000 individual laboratory values, and >700 antiviral administration records. In total, more than 2,500,000 characters were extracted from the EMR using this technique. Manual review of 10 randomly selected charts, found the data in perfect concordance with the EMR, a direct reflection of the fidelity of the parsing scripts. On average, an experienced user was able to enter three ICD-10 codes each minute, and six individual laboratory values per minute. At best, this same process would have taken at least 140 hours using a conventional chart review technique based upon the manual extraction times above.

This data was consolidated into a relational database that allows for direct visualization and statistical analysis of the many variables extracted (Poster #101).

Limitations:

This approach is an attempt to expeditiously retrieve information from the EMR by extracting only that which is presented to the user through the GUI. Ultimately, this technique still requires a level of manual chart review to extract the clinical data, which in some cases could still take hours to complete. While it is possible to access and directly extract these desired elements from the EMR, doing so requires the expertise of a clinical informatician who has the ability to specifically query the core database. For large scale analytics of tens of thousands of patients, or for real-time dashboard analytics of recurring clinical questions (i.e. how many grams of vancomycin were used this week), enlisting the help of a informatician is advised as those specific queries would be static.

Conclusions:

High-throughput data mining tools have the potential to improve the feasibility of studies dependent upon information stored in the EMR. When coupled with specific content knowledge, this approach can consolidate months of data collection into a day's task. While we have used this tool to extract and build a database of specific clinical outcome data for a cohort of patients with congenital CMV, this technique is generalizable to many clinical questions and provides an expeditious means by which to extract such information from the EMR. It should be noted that this technique is by no means a replacement for a clinical informatician, however it does serve to lessen the burden on these resources and focus the investigative skills of the physician for extracting clinical information from the EMR.

Acknowledgments:

We would like to thank the Pediatric Infectious Diseases Program at Texas Children's Hospital a member of the Baylor College of Medicine Affiliated Hospitals for their support of Dr. Rochat during his fellowship training.

DIGITAL POSTER