



A White Paper On EHR Data Supporting Pharmacovigilance Studies

How EHR Data Supports Pharmacovigilance and Aids in Signal Detection

*Peter Bechtel, President/CEO
eCast Corporation
November 18, 2008*

Overview

It has been known for some time by most clinical research professionals that the study of drug safety, pharmaeepidemiology and pharmacovigilance can be greatly enhanced with high-quality clinical data from Electronic Health Record (EHR) systems. Several teaching hospitals, such as Cleveland Clinic and Johns Hopkins University have successfully expanded their research capabilities by using clinical data. There are a number of EHR platforms being used by physicians for purposes of storing their clinical information; however the key to successful research with this data lies in the ability to successfully extract, store and present the data.

Definitions

EHR – Electronic Health Record – a software program and database system that stores medical record information about a patient’s health

EMR – Electronic Medical Record – same as EHR

PMS – Practice Management System – a software program and database system that processes billing and scheduling information for physicians and hospitals

ICD9 – International Classification of Disease Version 9 – A system of codification for diseases and diagnoses

CPT4 – Common Procedural Terminology Version 4 – A system of codification for procedures and time for medical care

NDC – National Drug Code – A system of codification for medications

LOINC – Logical Observation Identifiers Names and Codes - A system of codification for laboratory and other clinical observations

Vitals – Measurements of a patient’s height, weight, body mass index (calculated), temperature, blood pressure, waist size and other patient-specific medical values

Problems – A listing of the patient’s problems. These may sometimes be codified by using ICD9 codes, but frequently are in the physician’s “own words.”

Immunizations – A list of the vaccines and shots received by the patient as well as those scheduled to be administered

Attachments – Paper documents that are scanned into the EHR or EMR. These documents may sometimes have the words on the documents indexed and searchable

History – (Hx) – Narrative notes about the history of a patient’s health, including family history, genetic history, social history, past medical history, etc

Risk – Information about a patient that quantifies the profile of risk of the onset of certain chronic disease such as diabetes, congestive heart failure, etc

NLP – Natural Language Processing – A process of breaking narrated or dictated text into component parts that are stored as discrete nomenclature values such as SNOMED, UMLS or MEDCIN.

Sources of Data

Location of Data – EHR data is typically stored on a server (client-server) within the medical clinic. Traditional systems that have been in use for the past ten years almost always use a locally-centered server in which to house the EHR data. Only in the last five years have centralized database systems, or application service provider (ASP) system begun to emerge whereby the data is stored centrally.

Without delving into the advantages of client-server systems or ASP systems over each other, what is important to discuss are the practical limitations inherent in collecting data from these two systems.

Client-Server – Server based systems are difficult to work with when it comes to extracting data. Physical limitations of proprietary operating systems, security enabled database systems (Oracle, SQL, Cache, etc) and sheer location of the servers (closets, secure data rooms, under desks) make client-server systems difficult at best. The obvious exception is where the data

covers a large number of physicians and is managed by a professional Information Technology team who can assist with the extraction of data. With other client-server systems, such as those covering 2-5 physicians, the sheer cost and complexity of first locating the data and second extracting the data renders these systems impractical for use in pharmacovigilance.

ASP – Centrally located database systems are by far the best systems with which to use to extract data for pharmacovigilance. The central location and ability to extract data across single to multiple providers makes these systems ideal for data extraction. Generally such systems are located in professionally maintained data centers so it is important to have the access and cooperation of the governing Information Technology staff that manages the servers and data center.

EHR – What parts are valuable – The complete medical record consists of an abundance of data segments. Some of the data are useful and practical for drug pharmacovigilance where other segments are impractical. Following is a discussion of each segment and its pertinent usefulness to the drug study equation.

Demographics – Critical data includes patient gender, the year of the date of birth (without either month or day for privacy reasons) and the city or first three digits of the patient's postal code. If race is captured, this is a very valuable source of data for pharmacovigilance.

Vitals – Height + Weight (rendering body mass index) are highly desirable. Blood pressure readings are also of primary importance.

Problems – Typically, EHR data can include "free-form" problems such as "Constipation" as written by the nurse or in the patient's own words. Such data is extremely hard to standardize unless it is partnered with an ICD9 code.

ICD9 – Very useful data, especially when entered as part of the EHR data capture. This coding system tends to be of extreme importance in pharmacovigilance.

CPT4 – Secondary data and not often captured by EHR. When captured it can be extremely valuable to determine what was ordered (labs, pathology, and radiology) and what procedures were administered.

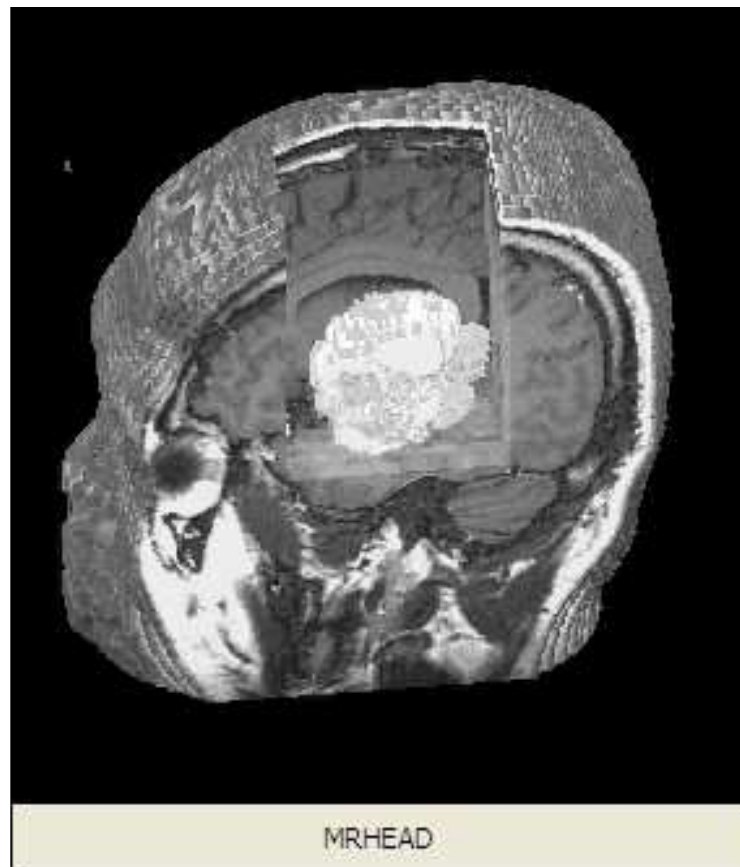
Lab Results – Extremely important in pharmacovigilance and readily available in huge quantities.

Lab Orders – Can be derived from HL-7 lab data (the “ORC” segment) and alternatively can be obtained from the EHR order entry system.

Pathology Results – Values that return “positive” and “negative” are most valuable to pharmacovigilance. Values that return verbose narrative are hard to incorporate into data analysis unless natural language processing (NLP) is employed against the results.

Radiology Results, Text – Very valuable data but only if NLP is employed against the results to turn the data into discrete nomenclature value points.

Radiology Results, Images – Not useful by themselves as there is no easy way to search images (see below) for any significant events or findings. Images are only useful when drilling down to study significant adverse events detected by signal detection.



Immunizations – The administration of vaccines and immunizations as a single binary event (given/not given) are the key factors of value in this data.

History, Text – Frequently, EHR data segments history apart from chart notes and visits. History is typically cumulative in that it is taken once and

“refreshed” with every patient visit. Text is the most common form of history in an EHR and it is of value only when processed through NLP to render discrete nomenclature values.

History, Structured Data from Templates – Many EHR systems use a point and click template structure to capture history. An example is the following:

History of Tobacco Use
Patient currently smokes cigarettes
Patient has never smoked
Patient smoked but quit XX years ago

In the example above, the EHR captures the values as input by the provider and stores the values as discrete values. If those discrete values can be captured and interpreted correctly, they can be tremendously valuable in pharmacovigilance. However, if the discrete values cannot be interpreted then the data would be considered inert.

Attachments – Attachments are typically stored as (1) images or (2) documents with searchable text. If the document is a PDF document that can be searched, then the “words” of the document can be processed through NLP and rendered into discrete nomenclature points which would be very valuable for pharmacovigilance. Attachments that are stored as images only (binary values) are inert for purposes of pharmacovigilance and signal detection.

Chart Notes, Text – Same rule as “History, Text”.

Chart Notes, Structured Data from Templates – Same consideration as “History, Structured Data from Templates.”

PMS – What parts are valuable - Frequently data is available from the PMS system within a medical practice. Physicians typically have been using PMS systems for as many as 30 years so data can be readily abundant. Here is a discussion about the data points available from PMS data:

Demographics – Similar to EHR demographics, gender, year of birth and city (or 3-digit postal code) are the most important components of Demographics from a PMS. Race is very rarely captured by front desk personnel and most PMS systems do not capture this data.

ICD9 – PMS Systems capture ICD9 codes as a regular method of billing to insurance companies, Medicare and Medicaid. However, it is not widely known that the ICD9 codes from a PMS system are neither always clinically

correct nor significant. Frequently, coding is exaggerated in order to obtain higher billing reimbursement. While not condoned by the government or insurance companies, adding ICD9 codes to patient visits can justify a higher level of service by the provider. Thus an elderly person, who is coming in for acid reflux problems, may be diagnosed with gastroenteritis and also lower back pain. In fact, most every human over the age of 50 probably has lower back pain of some kind, but when added to the gastroenteritis diagnosis, it can help justify a higher level of coding and service.

The other issue with medical billing codes is that some diagnoses are omitted deliberately so as to protect the patient from insurance company blacklisting. As an example, a patient with HIV typically may not want that diagnosis on his/her formal patient record with any insurance company. Therefore, to accommodate the patient, the provider will frequently mis-code the diagnosis to a lesser stigmatic condition.

It is for this reason, that when dealing with ONLY PMS ICD9 codes, the research team should be vigilant about data abnormalities and irregularities and treats the data accordingly.

CPT4 – All ambulatory PMS systems employ the CPT4 coding system for medical billing. These codes tell the insurance company what services and levels of service were performed by the provider. Generally speaking, these codes are highly accurate as they are a direct reflection of the work done by the provider.

Patient Intake Forms - Frequently researchers encounter providers who have bad quality data to no data at all, yet who still wish to provide research data.

Disease Risk Assessment

Instructions: Please mark one answer from each question or group. Do not mark outside the circle. Do not put any stray marks on this form.

Name: _____ Org Id: _____

Acct #: _____ Provider Id: _____

Personal Health Information

Ethnicity
 White African American Hispanic Other

Do you have a current diagnosis of, or have you ever been diagnosed with any of the following?

	Yes	No
Diabetes	<input type="radio"/>	<input type="radio"/>
Coronary Heart Disease	<input type="radio"/>	<input type="radio"/>
Stroke or TIA	<input type="radio"/>	<input type="radio"/>
Congestive Heart Failure	<input type="radio"/>	<input type="radio"/>
Other Cardiovascular Disease	<input type="radio"/>	<input type="radio"/>
Valve Disease or Heart Murmur	<input type="radio"/>	<input type="radio"/>
Left Ventricular Hypertrophy	<input type="radio"/>	<input type="radio"/>
Atrial Fibrillation	<input type="radio"/>	<input type="radio"/>

For Individuals Who Currently Smoke Cigarettes

	Yes	No
Has your birth mother or father, living or deceased ever been diagnosed with lung cancer?	<input type="radio"/>	<input type="radio"/>
Have you been diagnosed with lung cancer or COPD?	<input type="radio"/>	<input type="radio"/>
Have you been diagnosed with emphysema?	<input type="radio"/>	<input type="radio"/>
Have you been diagnosed with asthma?	<input type="radio"/>	<input type="radio"/>
Do you currently have asthma?	<input type="radio"/>	<input type="radio"/>
On average, how many cigarettes do you smoke daily?		
<10 <input type="radio"/> 10-20 <input type="radio"/> 21-30 <input type="radio"/> 31-40 <input type="radio"/> >40 <input type="radio"/>		
What is the combined number of years that you have smoked?		
<10 <input type="radio"/> 10-20 <input type="radio"/> 21-30 <input type="radio"/> 31-40 <input type="radio"/> >40 <input type="radio"/>		

Lifestyle Information

Do you currently smoke cigarettes?
 Yes No

Have you ever regularly smoked cigarettes?
 Yes No

Average times per week you exercise for at least 20 minutes:
 1 or Less 2-4 5 or More

While active, at what level of intensity do you exercise?
 Low Moderate High

Family History

Have any of your blood relatives, living or deceased, ever been told that they have any of the following?

	Yes	No
Diabetes At Any Age: Mother, Father, Sister, Brother, Daughter or Son	<input type="radio"/>	<input type="radio"/>
Coronary Heart Disease Before Age 55: Father or Brother	<input type="radio"/>	<input type="radio"/>
Coronary Heart Disease Before Age 65: Mother or Sister	<input type="radio"/>	<input type="radio"/>
Stroke Or TIA At Any Age: Mother, Father, Sister, Brother, Daughter or Son	<input type="radio"/>	<input type="radio"/>

Medications

	Yes	No
Do you take a high blood pressure medication?	<input type="radio"/>	<input type="radio"/>
On average, do you take the equivalent of half of an adult aspirin daily?	<input type="radio"/>	<input type="radio"/>
Do you take a cholesterol-lowering medication?	<input type="radio"/>	<input type="radio"/>

For Women Only

	Yes	No
Are you currently pregnant?	<input type="radio"/>	<input type="radio"/>
Have you ever had a live birth?	<input type="radio"/>	<input type="radio"/>
Were you ever diagnosed with gestational diabetes?	<input type="radio"/>	<input type="radio"/>
How many years since you were last diagnosed with gestational diabetes?		
<2 <input type="radio"/> 2-5 <input type="radio"/> 6-10 <input type="radio"/> 11-20 <input type="radio"/> >20 <input type="radio"/>		
Have you passed through menopause (either naturally or have had your ovaries removed)?	<input type="radio"/>	<input type="radio"/>
Are you currently using any form of hormone replacement therapy (after menopause only)?	<input type="radio"/>	<input type="radio"/>

Clinical Information (Office Use Only)

Height (Inches)	Pulse (Beats/Min)
Weight (lbs)	Systolic (mm/Hg)
Waist (Inches)	Diastolic (mm/Hg)

eCast has developed a system known as "SmartForms™" specifically for the purpose of collecting clinical data from physicians who have virtually no electronic clinical data. In truth, SmartForms was developed out of the frustration of knowing that a slim minority of physicians in the United States (purportedly 9-20%) actively employ an EHR (compared to European physicians who are in the range of 95-100%). With the eCast model of clinical research, data is critical for rapid enrollment, yet a significant number of investigators in the eCast site network have no EHR system and thus have no electronic clinical data.

The concept behind SmartForms is quite simple: Give the patient a form to fill out while he or she is waiting to see the doctor. Make the form simple (bubble fill-in) and readable (6th grade level) with large print and font. Provide instructions on the form so the patient knows how to fill in the bubbles.

As the forms are collected by the front desk attendants, they are stacked up and at the end of the day scanned into the clinical data repository (CDR) using a high-speed scanner. SmartForms also includes intelligent software that prompts the attendant if an answer is filled out incorrectly (e.g.: 2 answers to the same question) at which point the answer may be omitted or corrected by the attendant.

The SmartForms system is simple, foolproof and effective in collecting large amounts of clinical data quickly. It is used primarily as an alternative method of collecting data when there is no EHR available, but it can be used to supplement the EHR data as well.

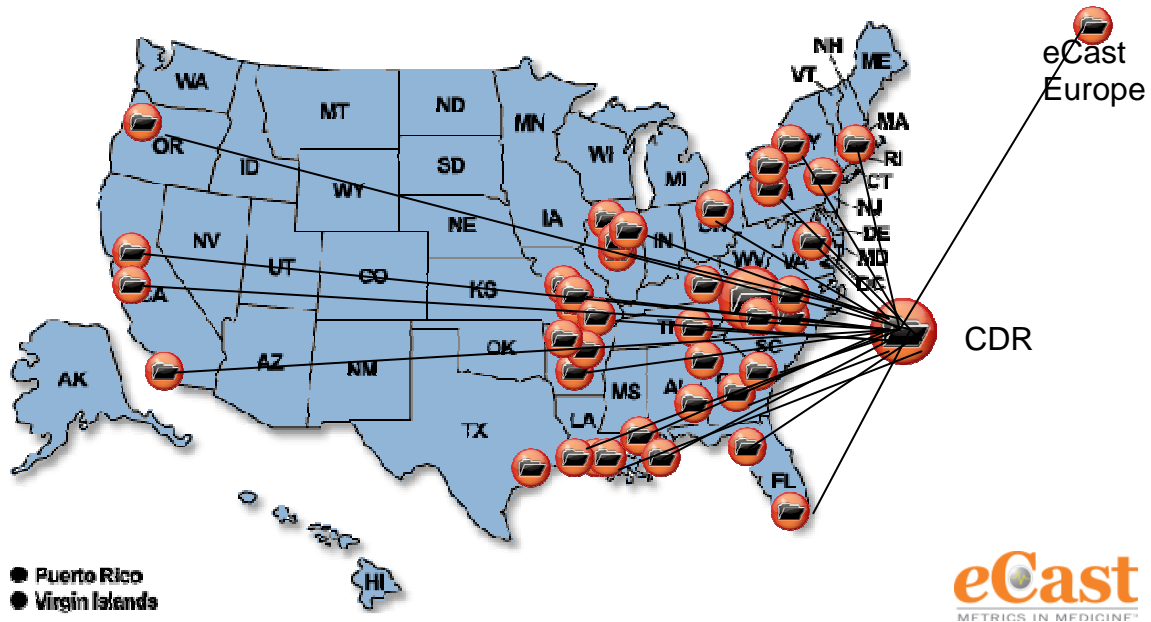
eCast has developed specialty specific forms for the provider as well. These forms zero in on the specific kinds of questions needed in clinical research and employ medical terminology familiar to physicians. With these forms, clinical data can be obtained that is highly useful to pharmacovigilance and takes just a few seconds for the physician as part of his or her routine.

To further automate SmartForms, eCast recently developed SmartTablet™ which essentially is SmartForms on the PC Tablet. Using SmartTablet, providers and nurses can walk around with a wireless tablet and click on the answers to the forms in question. This system removes the burden of printing and scanning forms and is actually preferred by physicians when they are the ones entering the data.

Whether SmartForms or SmartTablet is used, data for clinical research and pharmacovigilance can be readily obtained quickly and easily and with an extremely high degree of accuracy.

Clinical Data Repository

Data Source Locations



Given the various methods of collecting EHR and PMS data, eCast settled on a strategy of contracting with IPAs and PHOs around the nation in order to facilitate the collection of large quantities of clinical data. The Company knew that with a single IPA/PHO contract, massive amounts of data would be able to be obtained from the network of physicians IF eCast could render that same data back to the IPA/PHO in the form of their own data warehouse. The reason is simple: Most IPAs and PHOs have an aggressive requirement to share data among themselves because they all want to be either clinically integrated (meaning they can contract with insurance companies as a whole) or they want outcomes data for pay for performance (P4P) purposes.

With this in mind, eCast's strategy is to obtain contracts with provider networks and single tax-ID provider groups and in turn give them back the very data that is obtained from their providers. This is a tremendous benefit to the IPA/PHO because ordinarily that kind of data warehouse would cost them tens or hundreds of

thousands of dollars and take years to effect. With the eCast system, IPAs gain access to a data warehouse for their physicians at no charge and data is obtained wholesale from the network in these forms:

1. Lab data – Reference labs such as Quest and LabCorp will provide massive amounts of data from the IPA/PHO providers because they typically store it for up to 5 years. Therefore a single IPA/PHO can render tens of millions of lab results and hundreds of thousands of patient records.
2. Claims data – If the IPA/PHO has a contract with a payer, that payer typically will provide the claim data (which renders patient, ICD9 and CPT4 codes) back to the IPA/PHO and in turn can be imported into the eCast CDR.
3. Medications – Many IPA/PHOs have access to Pharmacy Benefit Management (PBM) companies whose responsibility is to manage the prescription writing and fulfillment on behalf of the network's physicians and patients. This data can be obtained through the IPA contracts and readily imported into the CDR. The final result of this data feed is normalized (on NDC code) medication data on patients.

Normalization

In building a CDR, the normalization of data codes is very critical. Normalization involves matching data to standard codes. An example of this would be medications. Since any CDR will receive medications from multiple data sources, one cannot use the text description of the medication in the query or analysis of the CDR. Instead, medications must be normalized to a standard code set. eCast uses the NDC (National Drug Code) standard for its medications. Other examples of normalization are the nomenclature codes for natural language processing (NLP). NLP is the process of converting narrative text (dictation as an example) into discrete data points. Three well-known nomenclature systems in use today are SNOMED™, UMLS™ and MEDCIN™.

NLP (Natural Language Processing)

The process of converting text from a History and Physical or Discharge Summary dictation note into discrete data values is of significant interest in Signal Detection. eCast partners with NLP International Corporation. In 2008, NLP International entered into an exclusive, worldwide licensing agreement with Columbia University for their NLP application - MedLEE™ (Medical Language Extraction and Encoding). NLP International licenses its "MedLEE™" product to eCast for conversion of chart notes into discrete data values known as "nomenclature." NLP International has chosen the UMLS™ nomenclature system but others such as SNOMED™ and MEDCIN™ can also be employed.

MedLEE turns unstructured, dictated medical narratives into easily retrievable accurate data to support multiple health care systems to enhance patient safety, quality assurance, diagnosis and prognosis support, billing and reimbursement administration. Because the translation is done on dictated or typed notes, the physician is not required to change work habits.

Developed at Columbia University during 20 years of research and extensive field testing, MedLEE is patented, highly modular and covers multiple medical domains. It is an extensively deployed, frequently referenced and widely published Natural Language Processor.

The significance of this cannot be understated insofar as providing the highest level of accuracy, granularity and thoroughness of the clinical history to the researcher. Nomenclature conversion provides a rich set of codes that are stored in the database along with the clinical record that can be searched and analyzed. As an example, compare the following text with its counterpart nomenclature coding:

Clinical information:

FEVER, RULE OUT HILAR ADENOPATHY.

impression:

MILD SCOLIOSIS OTHERWISE UNREMARKABLE FILM.

Description:

Frontal and lateral views of the chest demonstrate no evidence of mediastinal or hilar adenopathy. The lungs are clear. The heart size is normal. There is a mild scoliosis.

...for which MedLEE generated the following output without modifications to its processing engine or knowledge components ...

```
problem:fever
  idref>> 7
  parsemode>> mode1
  sectname>> report clinical information item
  sid>> 1
  code>> UMLS:C0015967_Fever
    idref>> [7]
problem:adenopathy
  bodyloc>> hilum
    idref>> 14
    code>> UMLS:C0929176_Hilum
      idref>> [14]
  certainty>> rule out
    idref>> 10
  idref>> 16
```

```

parsemode>> mode1
sectname>> report clinical information item
sid>> 1
code>> UMLS:C0149711_HILAR ADENOPATHY
idref>> [14,16]
problem:scoliosis
degree>> low degree
idref>> 25
idref>> 27
parsemode>> mode1
sectname>> report impression item
sid>> 2
procedure:film
certainty>> ignore
idref>> 29
idref>> 33
parsemode>> mode1
proceduredescr>> normal
idref>> 31
sectname>> report impression item
sid>> 2
code>> UMLS:C1705209_Film Device Component
idref>> [33]
procedure:view
bodyloc>> chest
idref>> 54
code>> UMLS:C0817096_Chest
idref>> [54]
idref>> 48
parsemode>> mode1
region>> front
idref>> 42
sectname>> report description item
sid>> 3
procedure:view
bodyloc>> chest
idref>> 54
code>> UMLS:C0817096_Chest
idref>> [54]
certainty>> high certainty
idref>> 44
idref>> 48
parsemode>> mode1
region>> lateral

```

```

    idref>> 46
    sectname>> report description item
    sid>> 3

problem:adenopathy
  bodyloc>> mediastinum
    idref>> 64
    code>> UMLS:C0025066_Mediastinum
    idref>> [64]
  certainty>> no
    idref>> 58
  idref>> 70
  parsemode>> mode1
  sectname>> report description item
  sid>> 3
  code>> UMLS:C1386346_mediastinal adenopathy
    idref>> [64,70]
problem:adenopathy
  bodyloc>> hilum
    idref>> 68
    code>> UMLS:C0929176_Hilum
    idref>> [68]
  certainty>> no
    idref>> 58
  idref>> 70
  parsemode>> mode1
  sectname>> report description item
  sid>> 3
  code>> UMLS:C0149711_HILAR ADENOPATHY
    idref>> [68,70]
finding:clear
  bodyloc>> lung
    idref>> 77
    code>> UMLS:C0024109_Lung
    idref>> [77]
  certainty>> high certainty
    idref>> 79
  idref>> 81
  parsemode>> mode1
  sectname>> report description item
  sid>> 4
normalfinding:normal
  bodymeas>> size of heart
    idref>> 88

```

```

code>> UMLS:C0744689_HEART SIZE
idref>> [88]
certainty>> high certainty
idref>> 92
idref>> 94
parsemode>> mode1
sectname>> report description item
sid>> 5
code>> UMLS:C0516868_HEART SIZE NORMAL
idref>> [88,94]
problem:scoliosis
degree>> low degree
idref>> 105
idref>> 107
parsemode>> mode1
sectname>> report description item
sid>> 6

```

The MedLEE NLP conversion system not only produces discrete UMLS nomenclature values, but also renders “confidence factors”. With any NLP system, a confidence factor must be employed to give the research team the ability to discriminate against “noise” and instead focus on highly reliable signals within the data.

Thus the conversion of chart notes to nomenclature gives the researcher a tremendous advantage when attempting to isolate the effects of a drug on hypertension as an example. Obviously, lab values and vitals can be used as a comparative factor, but to have subjective information and objective information included from the patient and the physician dramatically broadens the scope of the analysis. **When combined with all other discrete data elements such as medications, labs, vitals, immunizations and demographics, NLP nomenclature systems offer the highest level of searchable data possible given that the sources of data are wide and diverse.**

eCast and NLP International have partnered together to offer the conversion of all text documents from the clinical record into discrete nomenclature values that can be used by the researcher to isolate findings that may otherwise have been omitted or overlooked. Obviously this level of filtering and refinement can be immensely valuable in Signal Detection, Pharmacovigilance and Pharmaepidemiology.

A few notes about data from IPA/PHOs

1. All data obtained by eCast is stripped of patient health information (PHI) and 100% HIPAA compliant. The only indicator that is used is a master patient index (MPI) number. This number can only be used by the patient’s physician when he/she

needs the medical information from the CDR. Under these conditions, the physician is allowed to view the PHI because the patient is his or her patient and therefore the HIPAA security is not warranted. eCast provides a software program (eCast EMR or eCast Lite) with security controls that allows a physician to view this medical information for his or her patients. IPAs and PHOs do not have access to any PHI at any time. The only data available to IPAs and PHOs is de-identified "gross" data.

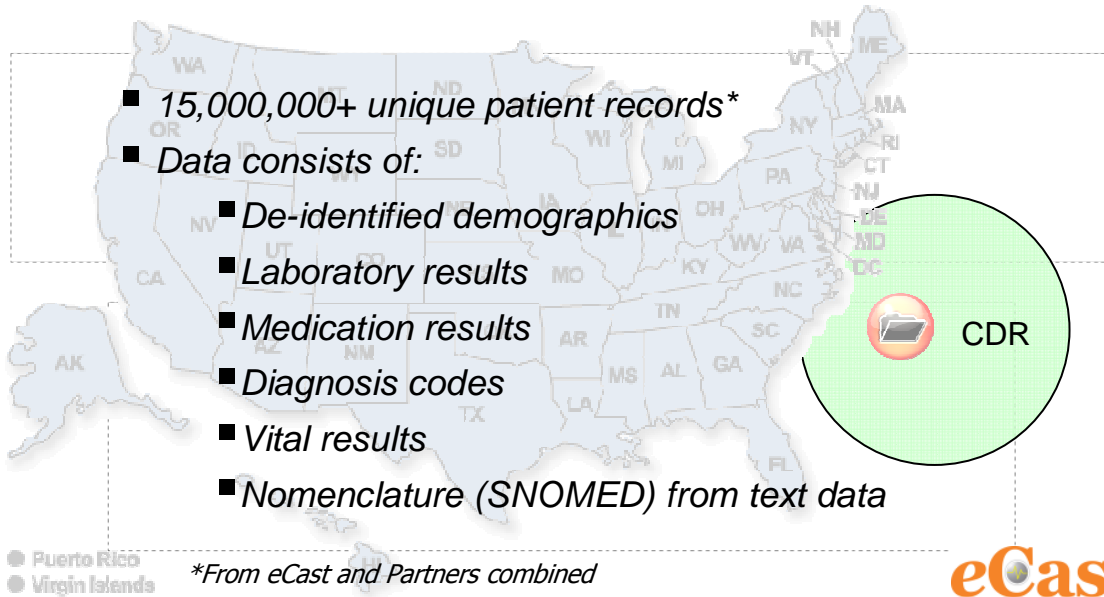
2. Lab data obtained on behalf of an IPA/PHO can be inherently "noisy" in that it has patient data that cannot easily be reconciled into a single master patient index. The reason is simple: Frequently patients are sent to a draw station for their blood work. The draw station personnel key the patient's data and the provider's name in every time they are seen because they usually do not keep the patient's medical record on file; therefore one patient (Jane Smith) being seen by one provider (Dr John Williams) may be keyed in under multiple combinations such as:

<u>Patient Name</u>	<u>Provider Name</u>
Jane Smith	John Williams, MD
Janie Smith	Dr. John Williams
J. Smith	Williams, John
Jane A. Smith	Williams, John, MD
Jayne Smithe	Dr J Williams

Therefore, researchers should be aware that there is an inherent "noise level" in lab data due to this factor. The actual lab results are typically perfect. It is the patient record that is generally the problem.

eCast has also partnered with other EHR companies who have large populations of physician data from tens of thousands of physician clients. This data is "virtually" attached to the eCast CDR. Translation: The partner's data resides in their own data center, but eCast has query rights to the data so that it becomes an extension of the core eCast CDR. Using this technology, eCast has extended its CDR by millions of data points.

Clinical Data Repository



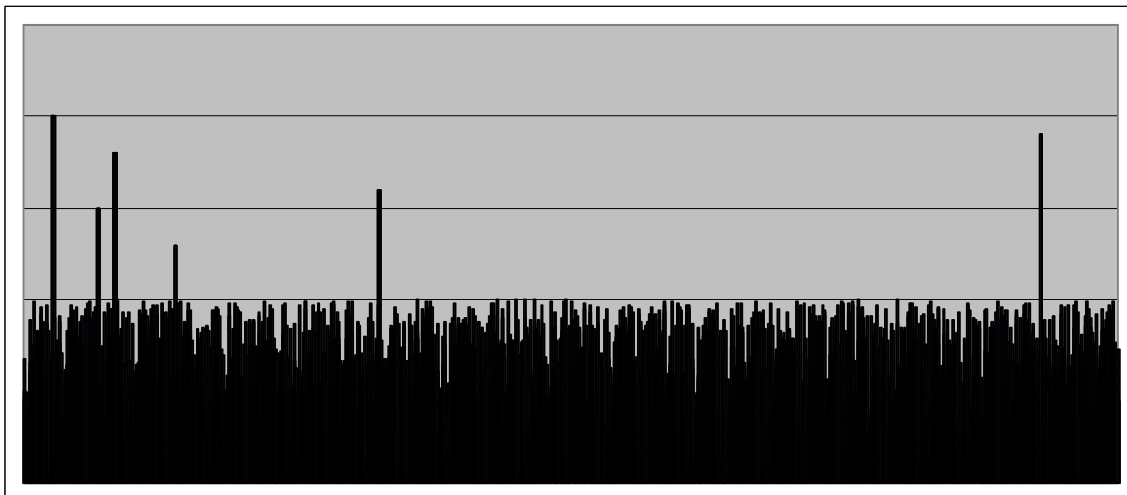
Pharmacovigilance and Signal Detection

We at eCast often use the motion picture "Contact" (starring Jodie Foster) as a parallel analogy of signal detection to make the point about how large quantities of clinical data can bring forth the "spikes" that are of interest to the pharmaceutical world. In "Contact", Jodie Foster patiently gathered trillions of bits of information from outer space, pumping the data through supercomputers in hopes of seeing the alien communication to earth in the form of a data "spike." Searching for drug adverse events has a similar methodology in that the researcher is pouring through millions of data elements to see patterns of lab values, blood pressure values, diagnoses or other indications.

As an example of this, assume that a drug company has developed a new drug that aids in the stiffness and pain of arthritis. Research was conducted to the satisfaction of the FDA and the drug was released to the market. Post-market Phase IV studies continued for several years.

The drug was used by millions of patients worldwide, but rare cases of CHF and Stroke were reported from random points. After researching these claims, it was found that the drug could cause heart failure and stroke in rare circumstances; even months after the patient discontinued the use of the drug.

In addition to the Phase IV study data, it is genuinely possible to study the drug using the theory of large numbers. That is, data that is sufficiently large will inherently contain aberrations and deviations (known as "spikes") that by themselves do not stand out, but when isolated among huge sets of data, display prominently. Illustrated below is a depiction of the "spikes" that emerge when huge data sets are analyzed. Jodie Foster in "Contact" captured huge arrays of data from outer space in the hopes of seeing patterns of "spikes" that could indicate an alien source. When analyzing CDR data, once the "spike" is identified, it can then be drilled down and analyzed and data from that analysis can be used to look through the entire data set for other patterns of similar data.



We at eCast believe that the critical mass of data for effective signal detection lies in the quantity of ten million patients or more. We must take into account that not every patient in the CDR has a complete set of data. In fact, only 30% typically have medication records, while 50% have lab data (according to the eCast CDR statistics). On the flip side of that is the fact that almost 95% of patients have diagnosis codes from either a PMS or EHR.

Value of the CDR

If drug researchers are trying to identify “spikes” from massive amounts of data, they must be acutely aware of the “value” of the CDR. In our definition, a CDR’s value is:

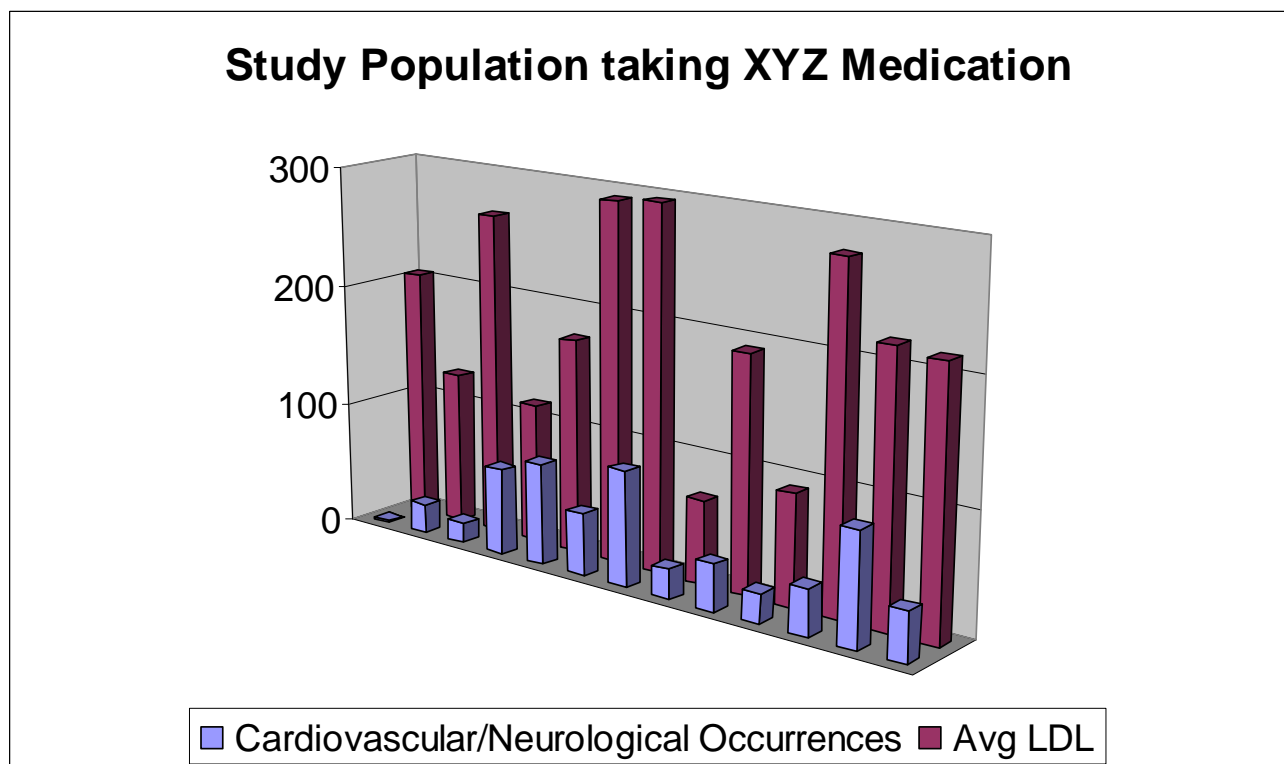
1. How many unique patient records exist?
2. How many patients have medication records and what is the average number of medications per patient?
3. How many patients have lab results and what is the average number of lab results per patient?
4. How many patients have vitals and what is the average number of vitals per patient?
5. How many patients have immunizations and what is the average number of immunizations per patient?
6. How many patients have history or text notes and to what extent are the nomenclature available?
7. Do all data points have consistent codification?
8. Do all data points have an accurate date of visit?
9. What is the reliability of the master patient index?

Example of CDR Data Analysis

In our fictional example of the arthritis drug “XYZ” that developed spurious cardiovascular events, researchers may have been able to see this anomaly retroactively by employing the following techniques:

1. Identify all patients who have a history of the drug XYZ in their medical record
2. Gather all lab values for those patients and filter those down to low-density lipoprotein (LDL)
3. Gather all ICD9 codes for those patients and combine all ICD9 codes that have any correlation to heart attack and/or stroke into a single set

4. Gather all nomenclature values for those patients and combine all codes that have any correlation to heart attack and/or stroke into a single set.
5. Establish Drug TimeLine: Plot out the gross data points for the medication utilization with an extrapolated "best guess" endpoint. In other words, if the patient were prescribed the drug for the first time on 02/15/2008 the data points would be plotted from that date forward through renewal dates and for some deterministic point thereafter (such as 90 days). The endpoint is an important consideration as adverse events can show themselves well after the last ingestion of the medication.
6. Against the Drug Timeline established by the medication utilization, plot out all adjacent lab values of interest (in our case, LDL values only).
7. Also against the Drug Timeline, plot out all occurrences of ICD9 codes and nomenclature codes that have been combined into a single query set. As an example, the patient's ICD9 codes for "Chest Pain", "Ischemia", "Myocardial Infarction", "Stroke", "TIA" and nomenclature for the same values combine together in a single group known as "Cardiovascular/Neurological Occurrences".



Extrapolation and Interpretation

Once CDR data is normalized and becomes available to the researcher, it can be "cubed", meaning analyzed against other potential important data points. As an example, the data above can be analyzed against blood pressure as well as Avg

LDL. Comparatively, such data could be stacked up against Avg LDL, Avg HDL and Total Cholesterol along with Triglycerides and possibly Lipoprotein-a, and Fibrinogen if those were available.

Summary

EHR Data can definitely aid researchers in their quest for pharmacovigilance and drug safety information. The data ideally should be highly organized, structured and contain normalized data values. In the best case data should be broad in content and deep in quantity. The research team should determine the “quality” of the data prior to the research study, and should employ the services of a qualified biostatistician to help render the most value out of the data.

Contact

If further information is required beyond this white paper, please contact the author:

Peter Bechtel, President
eCast Corporation
(919) 334-6300
pbechtel@ecastcorp.com