Is Your Healthcare Analytics System Missing the Gold in Your EHRs?

A Special Report for Healthcare Executives
Contents

Executive Summary 3

Checklist: Are You Mining All the Gold in Your EHRs? 4

What’s Being Used to Mine EHRs Today? 5
• Most Healthcare Systems are Quite Limited 5
• Valuable Information Found in Unstructured Data 5
• The Limits of Natural Language Processing (NLP) 6
• Using a Robust Taxonomy Yields Better NLP 7
• Extracting Gold from EHRs 9
• No Need to Rip-and-Replace 9

What Healthcare Organizations Need 10

What to Look for in an Ideal Solution 10

Introducing HealthData Engine 11

Conclusions 15

About Healthline 16

References 17
Executive Summary

The HITECH Act, ACA, and Meaningful Use have led to an explosion of electronic patient data. Successful healthcare organizations need to mine this electronic data for insights to help improve outcomes and manage costs.

Yet up to 80 percent of the data in Electronic Health Records (EHRs) is unstructured, free-form text in physician notes, patient histories, problem lists, hospital admission and discharge summaries, scanned documents, and so on.

Most healthcare analytics systems require data coded into the structured fields of an EHR, so they cannot access unstructured data. The result is that valuable clinical and non-clinical data is not being used, and healthcare organizations are missing important insights.

New technologies can help healthcare institutions extract more of the “gold” buried in their EHRs and other HIT systems. These technologies enable organizations to:

• Better identify high-risk/high-cost patients
• Optimize care delivery to reduce costs and improve outcomes
• Improve quality ratings
• Generate higher reimbursements
• Minimize readmissions
• Lower risks of penalties

To make the most of their EHRs, population health analytics, and other HIT investments, healthcare organizations should seek out these critical capabilities:

• Access to unstructured data in EHRs
• Advanced natural language processing (NLP)
• Comprehensive healthcare-specific taxonomy
• Fine-grained risk stratification
• Smooth integration with existing HIT systems

This white paper explores these capabilities in more detail and introduces a platform designed to provide all these capabilities: the HealthData Engine from Healthline.
Checklist: Are You Mining All the Gold in Your EHR?

There is gold in your EHRs. But just like gold ore in the ground, it's valuable—but challenging and costly to mine.

And if your organization is leaving much of that gold behind, you’re risking your clinical and financial outcomes.

How does your organization rate in mining value from your EHR?

If you answer “No” to any of the following questions, it may be time to consider adding more powerful healthcare analytics to your organization.

**Figure 1: Checklist for Mining EHRs**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Can you access unstructured data in your EHRs?</td>
<td>Y</td>
</tr>
<tr>
<td>2. Do you spend unacceptable amounts of time and labor combing through patient records manually to extract quality metrics, codes, and other clinical factors?</td>
<td>Y</td>
</tr>
<tr>
<td>3. Do you get real-time insights from your EHRs that are actionable at the point of care?</td>
<td>Y</td>
</tr>
<tr>
<td>4. Can you integrate these insights back into your EHRs?</td>
<td>Y</td>
</tr>
<tr>
<td>5. Can you integrate these insights with your clinical and administrative workflows?</td>
<td>Y</td>
</tr>
</tbody>
</table>
What’s Being Used to Mine EHRs Today?

How much analysis of EHR data is being done today? For many organizations, the answer is very little.

**Most Healthcare Systems Are Quite Limited**

As shown in Figure 2, the typical data analytics systems used by many healthcare organizations only scratch the surface of the value in their EHRs.

These systems use data transmitted in HL7, CCD/CDA, or other formats. But this type of structured data makes up only 20-40 percent of the total data stored in an EHR.

Analytics systems have trouble processing the other 60-80 percent contained in unstructured data, such as free-form text. Thus, these systems miss much of the content in those records.

**Valuable Information Found in Unstructured Data**

Here are some examples of valuable information found in the unstructured text in EMRs from recent studies at University of North Carolina Health Care and Seaton Healthcare:

- **CRC screening (colon cancer) values** are present in structured data just 17 percent of the time, but are found in unstructured data 83 percent of the time.

- **Left ventricle ejection fraction values**, which measure volume of blood pumped by the heart and help evaluate heart failure patients, are found just two percent of the time in structured data. However, these are found nearly three-quarters of the time in unstructured data.

- **Smoking indicators** were found in both structured and unstructured data, but their accuracy was markedly different. Structured data values were accurate 65 percent of the time, while unstructured data values were 95 percent accurate.

- **Type of living arrangement** can be a valuable predictive factor for hospital readmissions. While living arrangement was virtually absent from structured data, found less than one percent of the time, it is reliably found in unstructured data 81 percent of the time.

Without a system that can mine unstructured data, healthcare organizations are missing key information with which to make decisions.
The Limits of Natural Language Processing (NLP)

Given the huge value in unstructured data, some “natural language processing” (NLP) systems have been developed to analyze textual EHR data. But while NLP is an essential component of effective healthcare analytics, it is not sufficient.

NLP uses software to process text and recognize “entities” such as a medical condition. When a doctor makes an explicit diagnosis, NLP software can use its controlled terminology to identify the diagnosed condition.

Consider a patient diagnosed with diabetes and documented in the physician notes. After recognizing diabetes, NLP software characterizes that entity along a number of dimensions, such as:

- **Subject-object**: Who has diabetes—the patient, or a family member?
- **Temporal**: Does the patient have diabetes now, or did she have it in the past?
- **Clinical**: Is it Type I or Type II? Does the person with diabetes take insulin?
- **Severity**: Is the diabetes mild or severe? Controlled or not controlled?
- **Negation**: Does the patient present have diabetes, or is the note saying the patient does not have diabetes?

As you can see, these NLP-based systems rely on explicit diagnosis or procedure codes—such as ICD-9 or CPT—to identify key predictive factors and to calculate patient risk.
Natural language processing is required but insufficient for most use cases, since it requires the factors (e.g., diagnosis) to be explicitly stated.

Without a system that can mine unstructured data and make inferences, healthcare organizations are severely limited in managing their patient populations.

**Using a Robust Taxonomy Yields Better NLP**

In many cases, valuable information is not explicitly stated in structured text.

For example, a recent study *(Identification of Undiagnosed Diabetic and Quality of Diabetic Care in the United States, Tim A Holt, Candace Gunnarson, Paul Cload, Susan Ross, Nov. 2014)* found that 38 percent of diabetics in a population did not have an explicit diagnosis of “diabetes” in their medical records; they were only identified by the presence of diabetic medications.

Healthcare organizations require a system that can infer the presence of factors based on evidence from terms and concepts, such as treatments, medications, symptoms, co-morbidities, and so on, along with rules for weighing the evidence.

This can be called “an inference engine.”

The capability to infer the presence of entities or factors requires robust knowledge of terms and concepts related to the entity, and an understanding of the context of those relationships.

This comprehensive web of knowledge is a hierarchical semantic taxonomy in which health concepts are associated with relevant attributes. In this taxonomy, concepts are also mapped to other related concepts to reveal links between and among medical terms and concepts, words and expressions.

Numerous healthcare terminologies, classifications, and coding systems have been developed and refined over the years, as shown in Table 1.

Linking these terminologies together and adding additional information about medical attributes and relationships creates a robust, comprehensive taxonomy.

This enables more powerful NLP (including the ability to infer) and supports more effective mining of unstructured data to produce important new insights.
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Name</th>
<th>Publisher</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCD</td>
<td>Continuity of Care Document</td>
<td>American National Standards Institute (ANSI)</td>
</tr>
<tr>
<td>CDA</td>
<td>Clinical Document Architecture</td>
<td>Health Level 7 International (HL7)</td>
</tr>
<tr>
<td>CPT-4</td>
<td>Current Procedure Terminology</td>
<td>AMA</td>
</tr>
<tr>
<td>HL7</td>
<td>Health Level 7</td>
<td>Originally the University of California at San Francisco (UCSF) Medical Center; now HL7</td>
</tr>
<tr>
<td>ICD-9</td>
<td>International Classification of Diseases</td>
<td>WHO</td>
</tr>
<tr>
<td>ICD-9-CM</td>
<td>International Classification of Diseases, Clinical Modification (used for reimbursement in U.S.)</td>
<td>U.S. National Center for Health Statistics</td>
</tr>
<tr>
<td>ICD-10</td>
<td>International Classification of Diseases</td>
<td>WHO</td>
</tr>
<tr>
<td>LOINC</td>
<td>Logical Observation Identifiers Names and Codes</td>
<td>Regenstrief Institute</td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject Headings</td>
<td>U.S. National Library of Medicine</td>
</tr>
<tr>
<td>NDC</td>
<td>National Drug Code</td>
<td>FDA</td>
</tr>
<tr>
<td>RxNORM</td>
<td>Normalized names for all drugs sold in the U.S. using Unified Medical Language System*</td>
<td>U.S. National Library of Medicine</td>
</tr>
<tr>
<td>SNOWMED CT</td>
<td>Systematized Nomenclature of Medicine Clinical Terms</td>
<td>International Health Terminology Standards Development Organization (IHTSDO)</td>
</tr>
</tbody>
</table>

Source: Healthline
Extracting Gold from EHRs

Any mining operation requires modern tools and systems. You can’t recover much gold using only a pick and a shovel. In the same way, extracting the gold from EHRs requires modern tools and systems.

In fact, healthcare organizations may require more power than their current analytics system can provide. Big Data is defined as a body of information “beyond the ability of typical database software tools to capture, store, manage, and analyze.”¹ It’s not surprising that mining more value from the Big Data in EHRs often requires more sophisticated technologies.

“Without the appropriate investments in contemporary healthcare IT that enables value-based care, existing systems will be pushed beyond the breaking point,” concludes a recent study from McKesson.²

McKinsey consultants echo the same sentiment, warning that “legacy systems and incompatible standards and formats” can defeat any attempts to gain from “more sophisticated analytics that create value from Big Data.”³

No Need to Rip-and-Replace

Organizations that have invested heavily in HIT must find a way to implement more analytics, without having to rip-and-replace their existing systems.

To help extract more insights from EHRs, what features should healthcare executives look for?

The rest of this white paper describes the key capabilities provided by an advanced healthcare analytics system, why these are important, and where to find them.
What Healthcare Organizations Need

Healthcare organizations need a system that can extract critical insights from the unstructured patient data in their EHRs and other systems.

Using a combination of NLP and a comprehensive healthcare-specific taxonomy, this system should be able to identify important clinical information that is explicitly present in the data, as well as important information which is not explicit but for which there is evidence.

This system should integrate seamlessly with existing IT systems, support interactive dashboards, and generate cohorts of patients sorted by risk factors. And it should be able to merge this data back into EHRs and other HIT systems.

What to Look for In an Ideal Solution

Here are five key capabilities that an advanced data analytics system must provide to truly mine more of the gold from EHRs and other HIT systems:

1. Access to unstructured data in EHRs
2. Advanced natural language processing (NLP)
3. Comprehensive healthcare-specific taxonomy
4. Fine-grained risk stratification
5. Smooth integration with existing HIT systems

Fortunately, there is an offering coming to market that provides all these capabilities: the HealthData Engine from Healthline.
Introducing HealthData Engine

The HealthData Engine (HDE) is a powerful analytics platform that includes Medically Guided Search™, a robust health taxonomy, proprietary natural language processing (NLP), and an advanced clinical rules database.

HDE is designed to fuel a range of healthcare analytics and population health applications looking to analyze and extract insights from disparate sources of patient data, both structured and unstructured.

Figure 3 shows the main features of the HDE, with input on the left, processing in the center, and output on the right.

On the left, the HDE aggregates both structured and unstructured data from EHRs, making this data accessible to the processing engine. The HDE can access and normalize the 60-80 percent of the EHR found in free-form notes, documents, and clinical narratives.

Figure 3: HealthData Engine from Healthline
Both structured and unstructured data are then fed into the HDE processing engine. This features advanced NLP software, powered by the industry’s most comprehensive taxonomy, which provides powerful mapping of terms and concepts (ontologies), commonplace patient descriptions, and clinical symptoms, diagnoses, and medications.

On the right, the HDE output includes easy-to-read dashboards, charts, and patient cohorts and lists. Flexible data connectors—both software APIs and XML files—link smoothly to existing HIT systems, and support healthcare applications existing today, and on the horizon for tomorrow.

Here are some more details on how the six key capabilities are implemented in the HDE from Healthline.

1. **Access to Unstructured Data in EHRs**
   The system can aggregate and normalize information from disparate sources, including EHRs, specialized departmental systems, and other systems used in administration, claims processing and so on.

   The HDE first identifies structured data, such as problem lists and medication codes. Then it digs into .DOC, .HTML, and .PDF files filled with patient narratives gathered from admissions, patient interviews, and physician notes. This is especially important for recognizing the many socio-economic factors identified by researchers as highly predictive of various risks.

   The system combines both structured and unstructured data to yield a more comprehensive view of the patient.

2. **Advanced Natural Language Processing (NLP)**
   The unique NLP in the HDE goes beyond recognizing explicit mentions of certain entities to truly understanding the unstructured data in free-form narrative text.

   To achieve this, the software draws on the system’s industry-leading health taxonomy (see point 3 below). The NLP software combs through unstructured parts of the EHR to identify un-coded factors, including psycho-social, socioeconomic, and environmental factors.

   Mining more value from EHRs provides clinicians and administrators with a more complete view of their patient populations. This supports more informed decisions and better patient outcomes at lower cost and with less risk.
3. Comprehensive Healthcare-specific Taxonomy

The HDE encompasses the terminologies, classifications, and coding systems listed in Table 1, as well as many proprietary extensions and enhancements developed by Healthline. For example, each medical concept contains a rich set of attributes and relationships to other concepts.

In all, this taxonomy covers a huge number of terms and synonyms, including:

- 700,000+ Medical Concepts
- 90,250+ Diseases
- 40,800+ Symptoms
- 69,750+ Procedures
- 26,800+ Drugs
- 2+ Million Semantic Relationships

This extremely comprehensive and robust healthcare taxonomy powers the NLP software, and enables it to recognize entities more effectively. Since the HDE taxonomy contains more concepts and cross-references, it can infer the presence of more entities based on evidence found in unstructured data.

4. Fine-Grained Risk Stratification

Researchers around the world continue to develop highly effective predictive models that identify causes, correlations, and risks for any number of chronic diseases and conditions.

Most of these models rely on codified conditions, procedures, and diagnosis. Thus, before any healthcare organization can apply one of these models, it needs all these clinical factors coded accurately. Since many conditions, medications, and diagnoses are found in unstructured data, organizations must be able to extract these before applying any predictive risk models.
As risk models continue to evolve, institutions need a flexible solution that can use different models to help stratify and analyze cohorts, and identify the most effective targeted interventions.

The HDE’s risk-stratification engine is designed so that these models can be swapped in and out, as desired. A model can be based on clinical studies, or on standard risk-adjustment models, such as the Hierarchical Condition Categories (HCC) model.

The integrated rules engine yields extremely fine-grained stratification to group and assess a patient population for the appropriate conditions.

These results help clinicians and administrators target specific patients who are most at-risk of certain conditions where medical interventions can generate the best outcomes, lower costs, and avoid penalties.

5. Smooth Integration with Existing HIT Systems

With the HDE’s visual dashboard and patient lists, clinicians and analysts can see a comprehensive view of their entire patient population, stratified by risk or grouped into risk cohorts.

The extracted insights and risk scores can be merged back into the EHR, or sent to the core transaction system to be accessed as part of a clinical or administrative workflow.

The HDE can be integrated with most major healthcare analytics systems and EHRs, using standardized file formats such as XML or CCD/C-CDA through web services or APIs. This enables teams to continue using essentially the same familiar tools and workflow to achieve better results.
Conclusions

Electronic health records contain valuable information, especially unstructured data in physician notes, patient histories, and hospital admission and discharge records.

Typical healthcare analytics systems can’t use this free-form text. That means most healthcare organizations are leaving 60-80 percent of the value in their EHRs untapped.

Using the right tools to dig for the gold in EHRs will help generate:

- Better health outcomes
- Lower costs
- Fewer readmissions
- Higher quality ratings from patients and regulatory bodies
- Higher reimbursements
- Less risk of any penalties

To achieve these results, organizations may need to upgrade their data analytics systems. They should make sure to seek these key capabilities:

- Access to unstructured data in EHRs
- Advanced natural language processing (NLP)
- Comprehensive healthcare-specific taxonomy
- Fine-grained risk stratification
- Smooth integration with existing HIT systems

The right set of tools and technologies can create better outcomes and minimize risk for patients, clinicians, and the entire organization.

To find out more about how your organization can mine more value from EHRs to improve outcomes and reduce risks, visit:  
About Healthline

Healthline provides intelligent health information and technology solutions that help healthcare organizations and everyday people make more informed healthcare decisions, improve outcomes, and reduce costs.

Powered by the world’s largest medical taxonomy platform, Healthline’s HealthData Engine, Health Search Solutions and Health Marketing Solutions leverage advanced concept-mapping technology to deliver accurate, actionable insights. Additionally, the company’s consumer website, Healthline.com, delivers relevant, timely health information, news, and resources to help consumers manage their health.

Healthline is currently used by more than 25 million consumers per month and some of healthcare’s largest brands, including AARP, Aetna, Elsevier, GE, IBM, Microsoft and UnitedHealth Group.

For more information, please visit corp.healthline.com and healthline.com.
References


