Paper 0831-2017

A Practical Guide to Healthcare Data:

Tips, traps and techniques

Gregory S. Nelson ThotWave Technologies, Chapel Hill, NC

Abstract

Healthcare is weird. Healthcare data is even more so. The digitization of healthcare data that describes the patient experience is a modern phenomenon with most healthcare organizations still in their infancy. While the business of healthcare is already a century old, most organizations have focused their efforts on the financial aspects of healthcare and not on stakeholder experience or clinical outcomes. Think of the workflow that you may have experienced such as scheduling an appointment through doctor visits, obtaining lab tests, or prescriptions for interventions such as surgery or physical therapy. As you traverse the modern healthcare system, we are left with a digital footprint of administrative, process, quality, epidemiological, financial, clinical, and outcome measures that range in size, cleanliness and usefulness.

Whether you are new to healthcare or are looking to advance your knowledge of healthcare data and the techniques used to analyze it, this paper will serve as a practical guide to understanding and utilizing healthcare. We will explore common methods for how we structure and access data, discuss common challenges such as aggregating data into episodes of care, reverse engineering real world events, and dealing with the myriad of unstructured data found in nursing notes. Finally, we will discuss the ethical uses of healthcare data and the limits of informed consent that is critically important for those of us in analytics.

Introduction

Healthcare seems to have distinct worlds: research, quality improvement, financial and economic outcomes, as well as clinical outcomes. We see these worlds manifested in topics presented at conferences and types of analysts across the healthcare ecosystem. There is a rich history of financial reporting in healthcare that predates most Electronic Health Record Systems (EHR).

In the 1970's we saw a push to move from paper to Electronic Medical Records (EMR). In the late 1980's and 1990's there was a shift to warehouse the data from the EMR while the 2000's introduced the age of clinical decision support. While legislative uncertainty dominates the executive boardroom, there is little doubt that we will continue to see an increasing focus on value based reimbursement models in healthcare. Over this span of time there was a corresponding shift in focus from operational efficiency to optimizing finances to patient outcomes.

In the modern era of Big Data it is no longer sufficient to view healthcare as separate worlds as we realize the most benefit from our analytic efforts when we integrate data across the enterprise.

Healthcare is different

Earlier we said "Healthcare is weird. Healthcare data is even more so." It is noteworthy that healthcare has adopted several lessons from other industries, including:

- **Aviation –** the adoption of standard checklists for procedures
- **Manufacturing –** using Lean and Six Sigma techniques to managing and improve quality
- **Banking** – managing risk and improving security innovations through methods such as electronic transfer of data through real time messaging and the adoption of standards
- **Restaurant/ Hospitality –** striving for a consistent, highly predictable product at an affordable cost

The digitization of healthcare data that describes the patient experience is a modern phenomenon with most healthcare organizations still in their infancy. While the business of healthcare is already a century old, most organizations have focused their efforts on the financial aspects of healthcare and not on stakeholder experience or clinical outcomes. Think of the workflow that you may have experienced such as scheduling an appointment through doctor visits, obtaining lab tests, or prescriptions for interventions such as surgery or physical therapy. As you traverse the modern healthcare system, we are left with a digital footprint of administrative, process, quality, epidemiological, financial, clinical, and outcome measures that range in size, cleanliness and usefulness.

This digital footprint is not only unique because we, as patients, are unique. It is complex because healthcare is complex. These complexities have implications for us as "analysts" and include:

- There is no standard patient identifier across healthcare systems in the United States.

- Every implementation of standard EHR software packages are not standard (e.g., Epic's implementation at one hospital may not look like any other hospital.)
- While care delivery can be standardized, every patient has a unique condition, medical history, gene make-up, social support and behavioral profile.
- Comparing products across health services and health insurance products are nuanced, differentiated, and complex.
- Marketing dollars are spent marketing components of the healthcare experience to different buyers. For example, supplies are marketed to physicians and Supply Chain Officers (SCO's); prescriptions to pharmacists and physicians; experience to the consumer (patient).
- Payments and reimbursements in healthcare is strangely complex. The buyer is most often not the payer and the price list (charge master) is rarely ever used.
- Healthcare is delivered by lots of players that's make up a single "episode of care".
- A poor healthcare experience can result in people dying.

Historically, analysis in healthcare has been focused on the easy stuff – financial reporting and analysis. From there, people started to look at supply chain – since it was easy to get much of that data out of the financial systems. At present the economic shift from fee for service to pay for performance requires that we leverage all data – financial, operational, and clinical – and form a comprehensive view of how healthcare is is managed and delivered.

Data in Healthcare

To fully appreciate the complexity of data found in healthcare, we need to first describe the cacophony of data generated during the regular course of patient care and how that relates to the entire population of healthcare.

In the table below, we depict the typical clinical encounter and the various data generated and the systems used to support data:

Category	Type of Data	Data Source	Description
Patient Care Delivery	Medication Allergies Demographics Encounter Diagnoses Procedure	EHR Patient Registries Pharmacy Medical Imaging Clinical Decision Support	• EHR/EMR o Data collected by allied healthcare providers to provide diagnosis and treatment as part of clinical care. o Include a variety of data including patient demographics, clinical diagnoses, (problem lists), narrative text notes (e.g. clinic or inpatient notes), electronic reports of procedures or tests, laboratory data, vital sign data, medication data, and order/entry data • Imaging o Images and related electronic data from medical imaging procedures such as ultrasonography (including echocardiography), CT, MRI, PET, angiography, etc.

Category	Type of Data	Data Source	Description
	Diagnostics (ordered) Diagnostics (results) Symptoms Scheduling/ Registration/ ADT		• Patient Registries: o Systematic collection or capture from EDC or EHR of data with the use of standard data elements and definitions o Used to measure quality of care, provide quality benchmarks, and conduct clinical research • Externally Reported Data o Secondary uses include the surveillance of disease incidence and prevalence or regulatory reporting.
Biometric	BioMed Device	Device RFID Instrument	• Individual patient data reflecting physiology, such as vital signs or other physiological parameters (e.g. physical activity) • Data are increasingly available through remote monitoring of medical devices (e.g. implantable cardioverter–defibrillators) and/ or wearable technologies • These data may be reported directly by the patient by are clinically directed
Omics	Gene sequence SNPs Labs	Microarrays Labs-on-a-chip 23andMe.com	• A broad range of physiological laboratory tests and 'omic' data, including genomics, proteomics, and metabolomics • Indicate individual characteristics of patients that might be used to inform precision medicine
Administrative Data	Claims Billing Cost	Revenue Cycle systems Payers Supply Chain Management System General ledger Cost accounting	• Data collected as part of the routine administration of healthcare, for example reimbursement and contracting. • Billing data based on claims submission to payer. Typically related to utilization (visit, procedures, etc.) but may be based on episodes of care. • Secondary uses include the assessment of health outcomes and quality of care.
Patient Ecosystem (Context)	Social history Family history Lifestyle Socioeconomic Social network Consumer	Fitness memberships Grocery store purchases Credit Card purchases Mobile apps LinkedIn Facebook Instagram Census/ Zillow	• Increasingly the importance of patient context and environment is being considered not only in the care of an individual patient, but in population health management. • Examples include a cardiology patient may be required to have a home visit before surgery to determine whether their environment is conducive to recovery.

Category	Type of Data	Data Source	Description
Exogenous	Environment Weather	Climate (NOAA) HealthMap.org GIS Maps EPA Phone GPS Public Health databases	• While seemingly unrelated, external data sources such as weather can be important in managing the health of patients. • Geospatial variables can be used in analytic models to determine factors which may influence or moderate outcomes.
Patient Reported Outcomes	Sleep Exercise Diet Surveys	FitBit Surveys Diaries Mobile apps	• Patient survey data that can measure patient-reported outcomes, including patient health status (e.g. symptoms, functional status, and quality of life) and the care experience (e.g. patient satisfaction) • Patient-reported data can also inform 'patient-powered' research networks or provide feedback on medical therapeutics (e.g. reports of adverse effects) as well as surveys on patient satisfaction, care delivery, physician or facility rating • Patient managed medical devices such as FitBit • Patient diaries used in clinical trials or as part of care regimens • These data differ from clinically directed in that these are individually directed by the patient

Table 1: Typical Data Found in Healthcare

As we can see from above the variety and volume of data continue to outpace most organization's ability to make use of these digital assets. As a point of reference, here are a few stark facts regarding the amount of data we see in healthcare:

- patient monitoring equipment generates out an average of 1,000 readings per second or 88,400 readings per day
- It is estimates that nearly 5 billion patients worldwide will use remote health monitoring devices
- A single human genome requires around 200GB of raw storage
- A single health system such as Kaiser Permanente can generate the following in a year
    - 7 million online prescription refills
    - 21 million lab results are generated
    - 12 million secure email messages are sent
    - 36 million medical records
    - The average hospital generates about 665 terabytes of data with Kaiser sitting at 10 Petabytes

An organization's ability to transform source data into actionable analytics will be the key to survival in the world of value based healthcare. We know, for example that the value of data increases as it is combined with other data and that real value is created in analytics when we combine data sources to seek new insights

(Nelson, 2014). As we move from individual data sources to integrated analytic views, we can then begin use data to solve real business and clinical challenges.

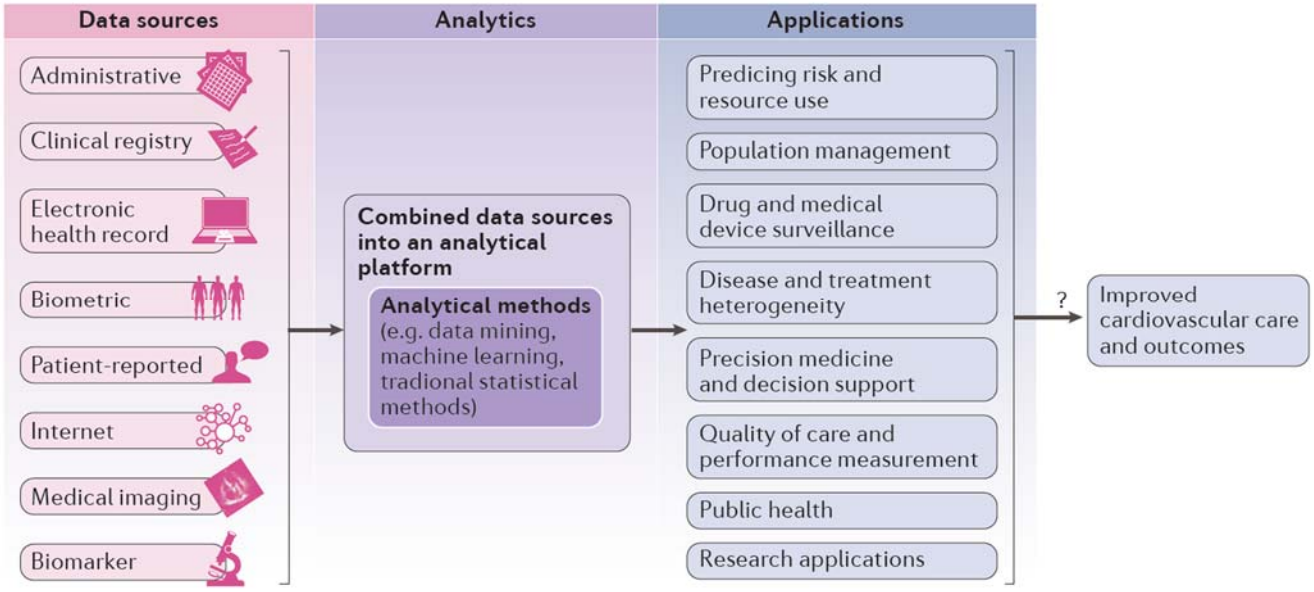


Figure 1: Examples of the inputs and outputs from a well-designed analytics platform (Rumsfeld, Joynt, & Maddox, 2016)

Understanding the Electronic Health Record

An Electronic Health Record (or EHR) is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all the key administrative clinical data relevant to that person's care under a provider, including demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data. and radiology reports. Depending on the breadth of an EHR, it can also include practice management and enterprise resource planning functions as to schedule patients and direct the flow of supplies.

A listing of the modules for the Epic EHR System can be found here: https://learnhealthtech.com/epic-systems-modules/. The EHR automates access to information and has the potential to streamline the clinician's workflow. The EHR also can support other care-related activities directly or indirectly through various interfaces, including scheduling, evidence-based decision support, quality management, and outcomes reporting.

EHRs are the next step in the continued progress of healthcare that can strengthen the relationship between patients and clinicians. The data—and the timeliness and availability of it—will enable providers to make better decisions and provide better care.

For example, the EHR can improve patient care by:

- Reducing the incidence of medical error by improving the accuracy and clarity of medical records.

- Making the health information available, reducing duplication of tests, reducing delays in treatment, and making patients better informed to take better decisions.
- Reducing medical error by improving the accuracy and clarity of medical records.

The early EHR era was dominated by a 'best of breed' approach to healthcare informatics. In considering all the different components that could go into the tool, healthcare organizations would choose the best vendor that suited each niche and sometimes would even develop their own applications. The problem with this approach was that there was no continuity among workflow of the different applications and often the resulting data was not housed in the same production database.

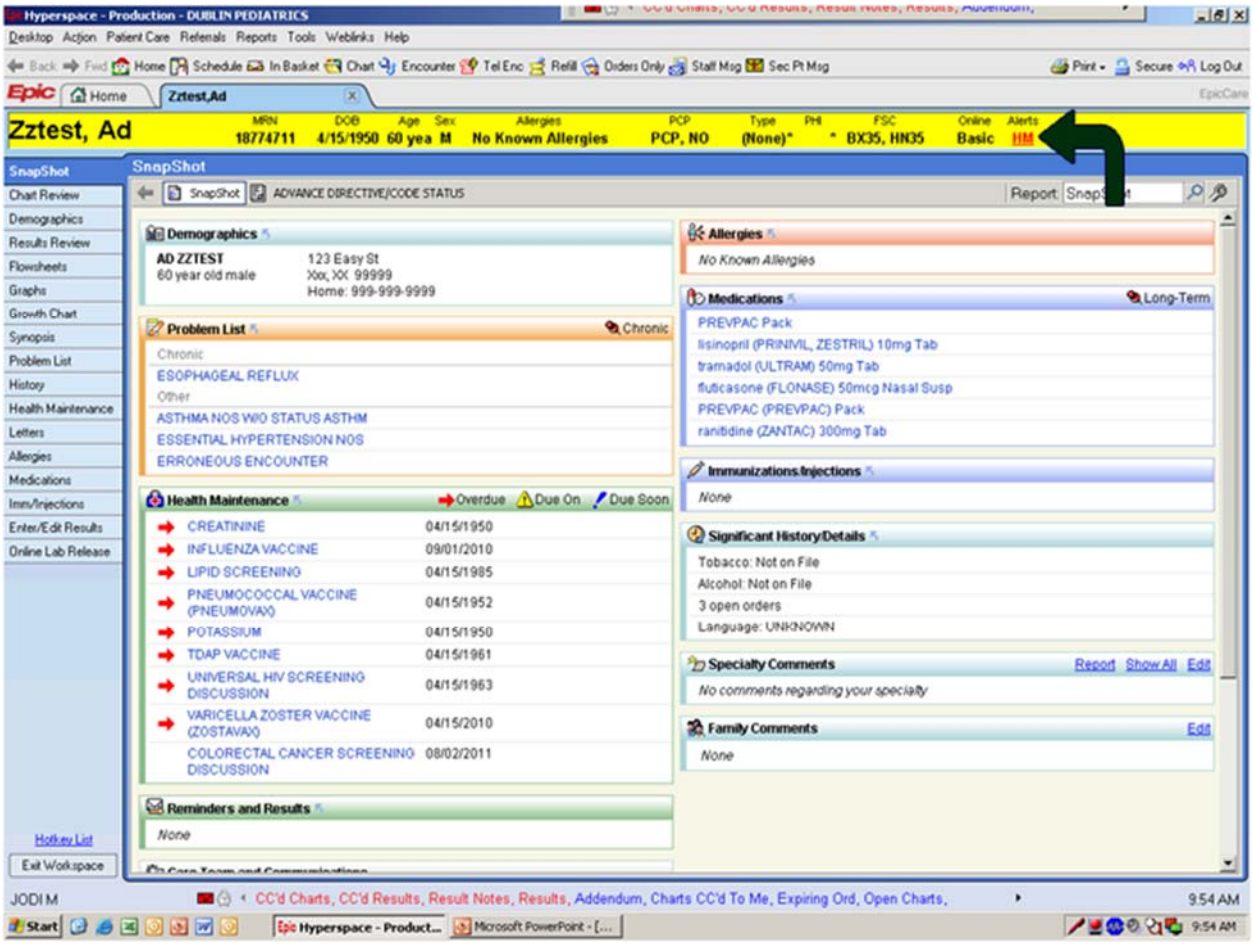


Figure 2: Example user interface for a patient record in Epic's EHR

Many tasks (such as entering allergies) would have to be done often for each application thus creating inefficiency and room for patient safety errors. Moving data between applications was also difficult, and some organizations still relied on paper. Other organizations were fully paper because of this, and never transitioned to an electronic system at all.

More recently, with the advent of the Federal Meaningful Use incentive program which we will talk about shortly, provider organizations have been seeking a single sign-on solution, that is, an integrated system that

shares information seamlessly across all care sites. This means that a clinician should only have to login once to one system to get what they need. Most commonly this is being done by choosing a single vendor product.

In this context, the EHR is becoming the centralized point of contact for all the different systems needed to monitor, advise, and coordinate care. As provider customers pressure EHR vendors to evolve product lines to keep up with government mandates, data sharing between different EHRs at different care organizations, often referred to as “interoperability” has become both increasingly critical and difficult.

WHAT’S IN AN EHR

An EHR is an operational system. It is designed to manage data primarily for one patient interaction at a time. The basic unit of analysis in any EHR is the encounter. The quote below describes the essence of an encounter.

> *When a patient seeks help from a health care provider such as a physician, nurse practitioner, or physician assistant, we describe this as an "encounter". There is a certain pattern, even a choreography, to that encounter, framed by patient expectations and the clinician's training. Visits can range from routine to emergent, and include routines and ceremonies and expectations. (Miller, 1992)*

Note that there are various clinical encounter types, the most common include:

1. **Routine visit –** these include simple, single visits in which a traditional fee-for-service model dominates. Examples may include an annual visit, medication checkup or diagnosis of a sign or symptom.
2. **Hospital admission** – Planned or unplanned admission into a hospital for treatment
3. **Virtual visits** - Increasingly established patients can interact with their provider over the telephone or via a telemedicine application in lieu of being seen in person
4. **Episode of care -** encounters that are linked to a common disease state. An episode of care consists of all clinically related services for one patient for a discrete diagnostic condition from the onset of symptoms until treatment is complete. For example, the visits associated with a normal pregnancy may constitute treatment within a maternity episode of care. Other examples include hip and knee replacement, diabetes and heart value replacement.
5. **Dramas –** Miller (1992) defined a third type of clinical encounter that he referred to as "dramas" which were a series of visits concerning situations of conflict and emotion and included psychosocial problems. These are often attributed to complex chronic disease states that have co-morbidities (e.g., diabetes and CHF.) Note: this is not well understood or implemented in current EHR systems.

With any encounter, there is a determination of the presenting concern, symptom or trigger, for a patient to be seen. Depending on the type of encounter you may see different types of data. The entity relationship diagram below illustrates some of the common data elements found in an EHR.

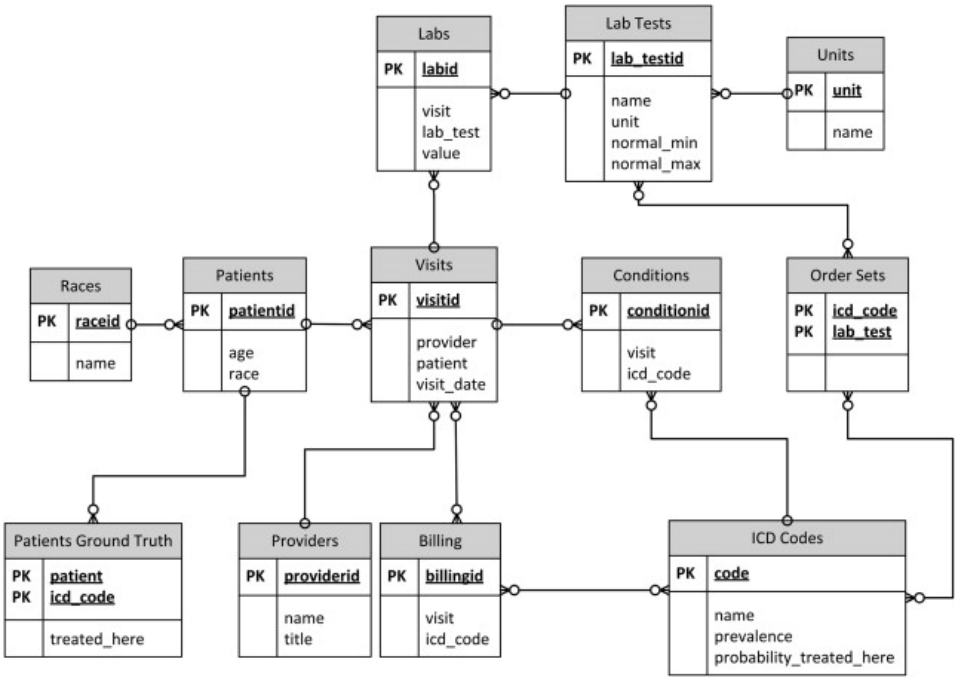


Figure 3: Sample entity relationship diagram for a healthcare organization

For an excellent example of a mature data warehouse in healthcare, take a look at the University of Pennsylvania Health System Data Warehouse. Information on Penn Medicine can be found at http://www.med.upenn.edu/dac/penn-data-store-warehouse.html and includes a data dictionary which can be downloaded. The table below illustrates the types of data captured in their enterprise data warehouse.

Patient demographics	Encounter/Visit details
Patient bed movements	Coded diagnosis
Coded procedures	DRGs
Medical history	Allergies
Medication administrations	Lab results
Microbiology	Blood Bank
Pathology	Vital Signs
Surgery	Anesthesia

Table 2: Data domains for Penn Medicine's Data Store

Analytic Challenges in Healthcare

Given what we learned above about the variety of data (e.g., images, video, real-time, voice, telemetry, unstructured notes) and the relative infancy as an industry in managing, governing and analyzing data, it is no

wonder that we see struggles to make headway in the use of advanced analytics in healthcare. The diagram below highlights some of the challenges that we face in analytics.

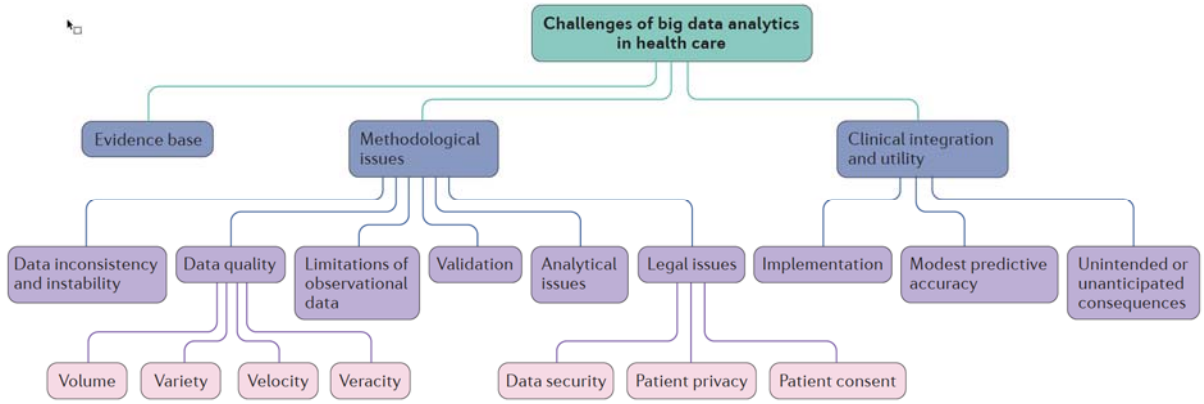


Figure 4: Challenges for Big Data applications (Rumsfeld et al., 2016)

There are several factors that contribute to these challenges and include methodological issues, philosophical, ethical, legal and overall utility/ usefulness. In the table below, we summarize some of the key issues related to common healthcare initiatives.

Initiative	Description	Challenges
Population Health Management	There are two primary goals of population health (1) increasing or improving the overall health and (2) reducing disparities within the population. In order to meet those goals, you must have an accurate picture of the following: 1. An understanding of the context of service (provider, location, service, patient) 2. The cost and utilization of care for all patients and services 3. Standard measures of quality and health outcomes (including mortality and complications) 4. External factors which influence healthcare disparity	Data from a variety of sources must be combined to accurately understand the services provided to the patient population and what is working. These include: • Insurance Administration Data (Claims, Pharmacy, Enrollment Data) • Clinical Data (Electronic Medical Records, Lab Results, Registries) • Medical Management Data (Health Risk Assessment, Authorization, Disease/Case Management Data) • Provider Administrative Data (Physician Practice Management, Hospital Billing, Admissions Discharge, and Transfers Data) • Public Data (State Discharge Datasets, Immunization Registries) • Demographic and Social Media Data (Lifestyle, Interest)
Predicting Patient Readmission	Since preventable readmissions in the healthcare setting are a major driver for CMS, hospitals are focused on patients that are at risk	In addition to the myriad of data sources needed to predict potential readmission (see above for desired data), factors exist outside of realm of control for most healthcare

Initiative	Description	Challenges
	for readmit. To achieve this, a healthcare system must have: 1. An understanding of the context of service (provider, location, service, patient) 2. Longitudinal data on all patients at the time of admission including a complete history of inpatient, outpatient, pharmacy, ambulatory and prior use of services as well as comorbidities and polypharmacy 3. Data on the comparative effectiveness of "impactable" factors that have been shown to reduce admissions 4. Social factors such as number of address changes, census tract socioeconomic status, history of cocaine use, and marital status	organizations. These include psycho-social, demographic or socio-economic factors. This becomes not only a data challenge but a system challenge as to how to affect and influence patient behavior. Finally, there appears to be growing evidence that traditional models are insufficient to demonstrate effective reduction of readmissions. See for example: • http://www.sciencedirect.com/science/article/pii/S1532046415000969 • http://jamanetwork.com/journals/jama/fullarticle/1104511 • http://www.jabfm.org/content/29/1/50.full • https://www.advisory.com/daily-briefing/2013/03/26/the-seven-factors-that-could-predict-readmissions
Learning Health System	A Learning Health System is essentially a delivery system that acquires the ability to routinely study and improve themselves. Recently, the Agency for Healthcare Research and Quality (AHRQ) opened up a new RFI which outlined their interests in the LHS. Note many of these questions center around the organizations use of data.	In addition to properly managing a vast repository of clinical and administrative data, organizations must: • Collaborate both internally and externally • Establish governance strategies around the use of data and analytic results • Capture knowledge about what's working and what is not working • Implement a data quality programs to ensure continuous improvement is baked into the data processes • Learn how to operationalize analytics and incorporate lessons from real world projects • For more information on the LHS please see http://www.learninghealth.org

Table 3: Analytic Challenges in Fulfilling Major Healthcare Initiatives

You will note with all three of the example initiatives provided above, access to a wide variety of data allows for broader health analysis which can lead to new and earlier insights to help identify improvements, efficiencies and effectiveness of the healthcare delivery system.

The examples above present a case for a well-integrated data strategy across the enterprise. Unfortunately, we remain an industry with lots of data, very little information and even less knowledge. Regulatory reforms, along with pressure from payers to improve quality and reduce costs will, no doubt, continue to shape our analytic priorities for the foreseeable future.

Foundational Requirements for Enterprise Analytics

The healthcare industry realizes in principle the importance of analytics in enhancing care quality and economic viability, but the execution of data and analytics strategies tends to fall short for many organizations. The core problem here is not technology nor a lack of analysts able to use Big Data tools. The real issue is that the ability of analytic professionals to make an impact on furthering a health organization's mission is dependent upon the unique confluence of analytic readiness (i.e. culture) and analytic maturity at that organization. Without the foundational underpinnings of data governance, change management, and systems thinking, analytic projects at best can attain a pilot state. It may be worth noting that there exist challenges related to both data governance and change management that really need to be addressed. For example, see the recent news about IBM Watson's success (or lack thereof) at the University of Texas Health System. (*The University of Texas System Administration Special Review of Procurement Procedures Related to the M.D. Anderson Cancer Center Oncology Expert Advisor Project*, 2016) http://www.utsystem.edu/sites/utsfiles/documents/system-audit/ut-system-administration-special-review-procurement-procedures-related-utmdacc-oncology-expert-advis/ut-system-administration-special-review-procurement-procedures-related-utmdacc-oncology-expert-advis.pdf

The single most important requirement for analytics success is a strong data governance program with well-defined master data. Master data management comprises the processes, governance, policies, standards and tools that consistently define and manage the critical data of an organization to provide a single point of reference. This corporately-managed master data should minimally include the data domains of patient, clinical provider, and care site location such that the data values mean the same thing regardless of the business unit doing analysis or the data source used. A strong enterprise data governance council needs to designate data stewards that help define the business meaning of data and proactively investigate quality issues. Data stewards are essential collaborators in any analytics project as there are often important clinical workflow nuances that can shape data interpretation.

Managed data is very important in healthcare because it ensures that concepts such as length of stay, admission date, and hospital visit type, to name a few, are consistent across the healthcare organization. All healthcare analysts should investigate if there is a knowledge management clearinghouse at their organizations in advance of any analysis to ensure they are getting the enterprise-vetted data and applying the right definitions.

Ultimately the constellation of these process and practices culminate to define an organization's data strategy, which is an actionable, comprehensive vision for how the organization uses and gleans knowledge from data. The data strategy helps define what data streams have priority, how much integration should take place, the

selection of master data domains, and the governance model for usage, maintenance, and service level expectations of data. Without a clear data strategy in place, it can be highly difficult to define the organization's source of truth and many duplicative data definitions and initiatives tend to germinate among business silos.

Integrating Data

As we noted above, healthcare creates and ingests a wide variety of data of differing purposes, volume, and velocity. The purpose of data integration is to provide the data assets needed to ensure that analysts can explore business questions with minimal data management and cleaning. Ultimately, the goal of a health analyst is to use the integrated to data to recapitulate and aggregate the real-world events that contribute to a clinical or business outcome. Although resource-intensive, a strategy for effectively integrating healthcare data over time according to its end purpose can answer many tough questions in healthcare.

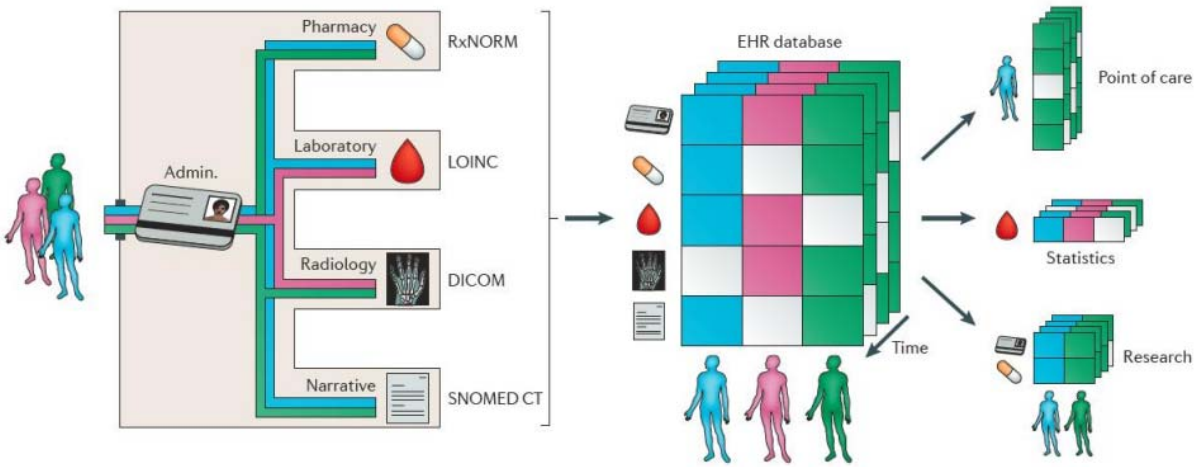
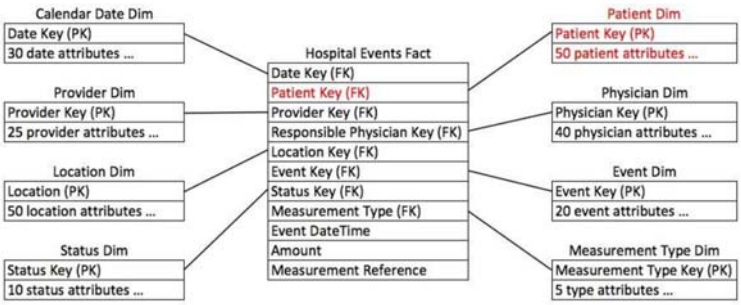
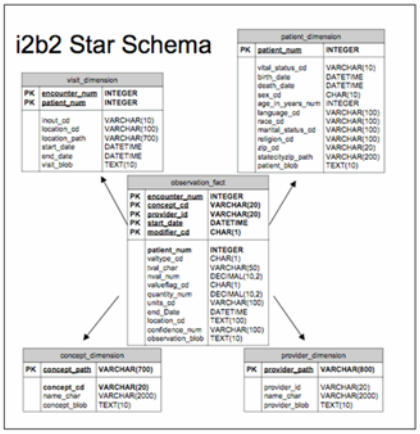


Figure 5: Mining Electronic Health Records ((Jensen, Jensen, & Brunak, 2012))

DATA STRUCTURES FOR INTEGRATION

Teams of technical analysts and data architects are common within health organizations and are tasked with the heavy lifting of data extraction, loading, cleaning, and transformation. As mentioned previously, much of the organizationally generated data comes from the EHR. Since the EHR is primarily designed for operational purposes (i.e., getting data for one patient in to the system or for looking up a single patient), they will often transform the operational data into a data model more suitable for analytics. In doing this, they are following a roadmap laid out by the organization's data strategy. The two diagrams below illustrate the organization of data in a traditional dimensional data model that describes a patient event during a hospital stay and a laboratory result (respectively.)





Across healthcare, we see numerous strategies for how organizations model and store data for analytics. These repositories usually include a variety of analysis-ready data marts where data is aggregated and presented for a specific business purpose. Regardless of the structure, gaining familiarity with the basic tenets of how data is represented is an essential part of becoming proficient with healthcare data.

Examples include:

- 3NF Data Warehouse (e.g., Epic Clarity, IBM, Teradata, Oracle, Duke's DEDUCE system)
- Dimensional Data Warehouse (e.g., i2b2, Epic Cogito/ Caboodle, John Hopkins Star Data Warehouse)
- Data Lake or Late-binding data warehouse (e.g., Health Catalyst, Mercy Hospital (Hadoop))

As the need to integrate data increases that doesn't have natural links, semantic models are starting to be used by organizations such as Montefiore Medical Center in New York (Sutner, 2016) to interrogate and find relatedness among disparate data sets. While relational databases represent data mostly linearly, semantic models permit spatial charting so that the relationship between data elements can be visualized. These models depend upon the distinct data streams being mapped to a variety of shared medical vocabularies, taxonomies, and ontologies. For example, even when two streams of data cannot be logically linked, it may be possible to look at subsets of the data and define how similar they are to each other use the ontology cross-reference. The data can be represented by the resource description framework (RDF) standard queried semantically using languages such as SPARQL. Graph algorithms are used to find unique relationships between different healthcare variables.

THE ELEMENT OF TIME

One sticking point in integrating and interpreting data is the concept of time. The wide variety of workflows and less-than-optimal business process in healthcare makes the concept of time particularly problematic for health data analysts. Often analysts need to follow patients through many different care encounters (see episodes, below) and create a rationale timeline.

A foundational challenge is that the expression of point in times comes via a date-timestamp which more often than not has a nuanced definition that requires exact understanding of the clinical workflow. Often analysts

make the mistake of assuming a date-time stamp is exactly when a clinical event happened. But timestamps can be created when one system writes a record to another, which may not occur in real time, and the timing of the actual clinical event may go unrecorded. For example, consider the fields in Table 3 where a timestamp could mean a variety of things depending upon a hospital's or clinic's unique workflow. The challenge here for analysts is that these definitions are very unlikely to be consistent across different healthcare systems and often even lack consistent meaning within a single organization. That is to say, these definitions can be highly site dependent.

Concept	Potential interpretations
Hospital admission (inpatient) date-time	• The time when a patient was registered at the hospital front desk • The time an order of admission was created by the admitting provider thus assigning a patient to a bed • The time the patient with an active admission order actually was placed in their bed, or 'roomed' • The time a patient was admitted to observation, which appears to be a hospital stay but may actually be an outpatient (e.g. clinic)-facing concept if the data point comes from a billing system • The time a patient presented to the emergency department prior to being admitted to the hospital
Clinic appointment (outpatient) date-time	• Patient check-in date-time at the clinic front desk • Scheduled appointment date-time • Date-time the patient's appointment was changed from 'scheduled' to 'arrived' • Date-time the patient's appointment was changed to 'completed' • Date-time the patient was placed into an examination room • Date-time when the provider saw the patient • Date-time when the provider closed the patient's associated medical record

Table 3: Potential definitions for two date-timestamps where a patient present to a clinical environment

One major confounder in analyses is that timestamps often trigger off of workflow actions, but if that action isn't mandatory (e.g. a staff member does not follow model workflow), then a timestamp may be missing or default to a system value. For example, in one major EHR system it is common to see appointments still listed as 'scheduled' even though they have been completed and even billed to payers.

Another challenge in is that knowledge management surrounding environmental changes is extremely poor and inconsistently documented. A healthcare setting is a living system that experiences system upgrades, remapping of staff responsibilities, workflow changes, and influences from the environmental such as flu season or weather. Not knowing these elements can leave important parameters out of analytic models. Some examples of these situations can include:

- Hospital workflow altered so that nurses, not doctors, order and interpret monitoring lab tests
- Clinics reducing appointment availability due to an EHR upgrade
- Learners unfamiliar with the tools, such as residents and interns, being added to a care environment
- Superbowl Sunday creating a dip in emergency department arrivals

The only satisfactory way to address this, short of a comprehensive enterprise change and knowledge management system, is to work closely with business analysts to thoroughly interview all stakeholders and develop a clear picture of the environment being modeled.

Unstructured Data

Unstructured data remains a huge analytic challenge for most industries, and healthcare is no different. There is a huge variety of data that are caught up in free text notes all throughout the clinical record that provide essential context and depth about the patient experience (Table 4). Despite being such a rich source, they also present a number of challenges to the health analyst:

- Misspellings, non-standard abbreviations, and grammar errors
- Extensive use of short phrases
- Lack of completeness depending on the author
- Different templates for different types of notes
- Note 'bloat' from copy-paste of another note or a separate structured source, such as labs, medications or vitals.

Data source	Features
Clinical notes	Highly available and detailed but difficult to process given stylistic difference between authors
Lab results	Molecular labs and those that correspond to more elaborate omic profiling methods will often have an unstructured explanatory note
Radiology and imaging results	A structured template is used but the note itself is free text
Social media	Insights into patient health possible but much volume of data with very little health-facing content
Medical literature	Highly variable text presentation and conclusions within may not stand the test of time
Patient satisfaction	Some structured data is available from well-known, vended surveys. There is a growing variety of reviews on social media and sites such as Yelp that have highly variable data
Clinical guidelines	Typically presented on a per-specialty basis which can ease the text mining process by limiting the search to certain medical vocabularies. Highly variable release schedule of new or updated guidelines

Table 4: Unstructured Data Found in Healthcare

Addressing unstructured data is very computationally intensive and usually starts with text mining algorithms that find features within the data followed by machine learning methods that seek to evaluate those features'

predictive capacity in modeling. In recent years, the healthcare industry has grown a number of products that support free text mining that are tailored specifically for unstructured health data. IBM Watson has been successful in using cognitive computing methods to better tailor treatments for a variety of cancers based on an individuals' clinical profile, molecular genetics profile, and insight from the medical literature. Natural Language Processing (NLP) is routinely used (Pakhomov, Buntrock, & Chute, 2006) as a first pass for assigning diagnosis codes to clinical text as to speed the process of billing, and has been packed into a product by M*Modal.

Due to its richness, unstructured data is sought out by researchers who may have the appropriate collaborations to put the advanced modeling and statistical methods in the hands of an experienced statistician or data scientist. But given it could contain almost anything, unstructured data does the run the risk of containing information that identifies the patient. Care must be taken in any associated study protocols to ensure reasonable safeguards are in place for handling the data even if it is believed that protected health information should not be present.

Episodes of Care

> *An episode of care (episode) is defined as the set of services provided to treat a clinical condition or procedure.*
>
> Source: Center for Medicare & Medicaid Services

Our current healthcare system functions based on the assumption that when a service is rendered, a bill is generated. Sometimes the payer is the patient, other times it is a private insurer or government entity (e.g., CMS, VA, DoD.) But in recent years there is increasing interest on the part of both payers and patients to change this and move from a fee-for-service to a fee-for-value model. The latter means that healthcare providers are would be paid mostly a fixed price, with some modifications, for a certain clinical episode of care. The exact formula and network of incentives varies dramatically by program, but value-based models include bundled payments, episode-of-care payments, and accountable care organizations, which are commonly referred to as 'alternative payment models'. When using one of these models, episodes are analytically constructed for each patient through the following process:

- Episode identification determines which episodes exist during the period under consideration and when those episodes begin and end.
- Service assignment ascribes each claim or service line to one or more episodes, and then determines whether each claim or service is typical for that type of episode or complication (or typical with complications).
- Cost allocation assigns the money for a claim or service to one or more episodes and then determines how the money should be distributed if the services are assigned to more than one episode.

Episode definitions do vary, and the constellation of payer arrangements will indicate what episode definitions are required for a clinical scenario.

While this concept makes a lot of sense, is difficult to implement for lots of reasons including the inexact science of treatment, presence of evolving therapies, and off-label uses of drugs. This is challenging because within an episode of care definition, the foundational assumption is that the amount of payment for a specific episode of care can be reliably set and variation addressed via regional adjustments to the payment model. As a result, this paradigm transfers the financial and legal risk nearly entirely to the healthcare provider even though there are so many factors that can influence patient well-being, including environment and social support, which lies out of the control of physicians. A second problem is that this shift also places the payer in a role such that they more directly influence medical decision-making. Patients are unlikely to approve treatments that insurance won't cover and a care episode typically defines an agreed upon order or operations for treatment. Some providers pejoratively term this as 'cookbook medicine' and view it as meddling in the patient-provider relationship.

Finally, we cannot underemphasize the implicit analytic challenges with bundled payments, particularly for those organizations whose data and analytic strategies have been shaped by the fee-for-service model. Analytics is essential to adequately protect healthcare providers from the risk. They must know their costs and be able to attribute discrete healthcare services and events to the bundle or episode as well as negotiate with savvy with payers during contracting. Even the largest and most prominent healthcare systems often don't have a handle on their actual costs. Since care is delivered by a multitude of players across organizational boundaries, transparency in costs, quality of care, and outcomes often remain elusive.

Ethical Uses of Data

People are trusting the history of their health in the hands of care organizations and expect that such intimate information about them be treated respectfully. All healthcare data usage is bound by the rules of HIPAA (the Health Insurance Portability and Accountability Act) and informed consent depending on the nature of the investigation. But this whole concept of ethics goes beyond just being compliant with patient privacy rules and HIPAA policy. Ethical data use means that all investigations are conducted while complying with the federal regulations for patient privacy. It requires daily vigilance into how data passes between stages in your analytic workflow. Even a simple oversight, such as using a non-HIPAA compliant storage site such as DropBox to share data with a coworker, puts your organization at risk.

Informed Consent

No matter the department you work, the instant that findings generated from healthcare data are generalized and sought to be shared outside of the business team – even at a vendor product user group meeting – you are in the realm of research and should seriously address the rules of informed consent. Informed consent is the process for getting permission before conducting a healthcare intervention on a person, and analyzing one's private healthcare data can be construed as an intervention. Challenges with Big Data and informed consent tend to come about when researchers want to combine huge repositories of data in novel ways as to ask questions of data that only very large populations can answer. Comparative effectiveness research, for

example, is a field that has the possibility of making great headway in understanding the differential cost-benefit of drugs and treatments over time. The speed of these discoveries, however, are constrained by the nature of the informed consent process where getting permission from every individual would be prohibitively expensive and time-consuming. Some health organizations, such as Vanderbilt University Medical Center (VUMC), take an opt-out approach where you are automatically opt-ed into big data investigations unless you sign specific paperwork mandating that your data may not be used.

Ethical Challenges in the Secondary Use of Big Data

Secondary use refers to the act of using data generated as a by-product of clinical care for a second purpose such as business process improvement, evaluation of compliance to clinical guidelines, or estimation of patient safety or care quality metrics. In the HIPAA world these investigations fall under the realm of 'quality improvement' meaning that the activity is exempt from notifying the patient that their data is being used for these purposes not related to their direct care. Big Data has challenged a lot of the traditional frameworks for quality improvement particularly when you consider the amount of data external to the healthcare environment that now can be combined with clinical data.

Imagine you are on an analytics team that uses a process where social media sites such as Twitter or Facebook were scraped for publicly-shared information about patients. This is common in the marketing and consumer research world, so it shouldn't be surprising that these data can be blended with clinical information to help answer questions. Even richer sources of data is available at a price from data brokers who curate massive proprietary databases that contain years' worth of inferred interests, demographics, household data, and purchasing behavior for individuals. In fact, one leader in the field processes more than 50 trillion data points yearly and has about 1,500 data points per person on most adults in the country. All it takes is a name, birthdate, and maybe email address to get information on an individual. But those using these strategies needs to take a lesson from design thinking and imagine the perspective of the patient.

- How would a patient feel about their buying habits informing a readmission risk model?
- When is it justified to use these data for care and when does it risk eroding trust between the patient and provider?
- Is that extra data truly useful and predictive?

These questions should give health data analysts pause, and we realize there are no easy answers. This is a conversation to have with your stakeholders in the organization about what is supported in the culture.

Summary

Few industries have a data environment as complicated and emotionally charged as healthcare. Yet despite its importance, the digitization of healthcare data that describes the patient experience is a modern phenomenon with most healthcare organizations still in their infancy. With this come the expected growing pains in developing tools that can collect and distribute data effectively so that it can be used for a variety of business

purposes. In this environment, analytics teams that have broad executive support are nearing an age of enlightenment. Whether they reach that point depends upon being able to resolve common traps inherent in the challenges of ethics, data interpretation, data integration, and maintaining alignment with the organizational data strategy.

Acknowledgements

I would like to thank Monica Horvath for her "thot-ful" review of this manuscript. Her intellect, curiosity and passion in support of the healthcare analytics work that we do continues to inspire me.

Biography

Greg Nelson, President and CEO, ThotWave Technologies, LLC.

Greg is a global healthcare and Business Intelligence (B.I.) executive with over two decades of experience and leadership in the field. Greg is a prolific writer and speaker interested in healthcare analytics and the strategic use of information technology.

Contact information

Your comments and questions are valued and encouraged. Contact the authors at:

Greg Nelson greg@thotwave.com http://www.thotwave.com

ThotWave Technologies, LLC

1289 Fordham Boulevard #241 Chapel Hill, NC 27514 (800) 584 2819

References

Jensen, P. B., Jensen, L. J., & Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet, 13*(6), 395-405. doi:10.1038/nrg3208

Miller, W. L. (1992). Routine, ceremony, or drama: an exploratory field study of the primary care clinical encounter. *J Fam Pract, 34*(3), 289-296.

Nelson, G. S. a. D., Lisa. (2014, March 23-26, 2014). *Modernizing Your Data Strategy: Understanding SAS® Solutions for Data Integration, Data Quality, Data Governance and Master Data Management.* Paper presented at the SAS Global Forum, Washington, DC.

Pakhomov, S. V., Buntrock, J. D., & Chute, C. G. (2006). Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J Am Med Inform Assoc, 13*(5), 516-525. doi:10.1197/jamia.M2077

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nat Rev Cardiol, 13*(6), 350-359. doi:10.1038/nrcardio.2016.42

Sutner, S. (2016). Semantic graph database underpins healthcare data lake. *Innovation Spotlight.* Retrieved from http://searchhealthit.techtarget.com/feature/Semantic-graph-database-underpins-healthcare-data-lake

The University of Texas System Administration Special Review of Procurement Procedures Related to the M.D. Anderson Cancer Center Oncology Expert Advisor Project. (2016). Retrieved from Austin, Texas: