Article type : Special Contribution

Abstract

For a variety of reasons including cheap computing, widespread adoption of electronic medical records, digitalization of imaging and biosignals, and rapid development of novel technologies, the amount of healthcare data being collected, recorded, and stored is increasing at an exponential rate. Yet despite these advances, methods for the valid, efficient, and ethical utilization of these data remain underdeveloped. Emergency care research, in particular, poses several unique challenges in this rapidly evolving field. A group of content experts was recently convened to identify research priorities related to barriers to the application of data science to emergency care research. These recommendations included: 1) Developing methods for cross-platform identification and linkage of patients; 2) Creating central, de-identified, open access databases; 3) Improving methodologies for visualization and analysis of intensively sampled data; 4) Developing methods to identify and standardize electronic medical record data quality; 5) Improving and utilizing natural language processing; 6) Developing and utilizing syndrome or complaint-based based taxonomies of disease; 7) Developing practical and ethical framework to leverage electronic systems for controlled trials; 8) Exploring technologies to help enable clinical trials in the emergency setting; and 9) Training emergency care clinicians in data science and data scientists in emergency care medicine. The background, rationale, and conclusions of these recommendations are included in the present manuscript.

Introduction

The promise of big data and data science to revolutionize many facets of society, including the practice of medicine, is a common refrain found in the medical literature, particularly within specialty and policy circles for the past several years.[1] These discussions have more recently begun to filter to healthcare providers, adding increasing relevance of the topic to clinicians, who are more likely to encounter such discussions While the definition of big data varies, it generally refers to some combination of increasing size and scope of data, including non-discrete "natural language" data fields, and the novel methods and tools to analyze such large and complex data sets. The potential for electronic data to improve patient care in the emergency department (ED) is particularly exciting, in light of the critical nature of many decisions made there. Electronic data capture to promote a better understanding of health and disease has always been part of the argument for the implementation of electronic health records (EHRs). Despite widespread deployment of such systems, there remains skepticism regarding the ability to actually deliver on such promised value.[2] Even with expansion of both infrastructure and computational power, significant barriers exist that limit advances to "learning healthcare systems". For a variety of reasons, including the acuity of medical conditions and a fragmented healthcare system, many of these barriers-- interoperability, data availability, and data islands-- globally relevant to novel healthcare data science are magnified when applied to emergency care.[3] Addressing the issues identified will allow for more streamlined use of and more valid conclusions resulting from research using these novel methods, and to provide new insights into pathophysiology, comparative effectiveness of clinical interventions, and clinical systems and operations. Failure to address the potential pitfalls will at best complicate the conduct of research, and at worst contribute to fundamentally flawed conclusions, with widespread consequences such as the development of flawed quality metrics or worthless interventions. For these reasons, identification of research and policy priorities for this field is particularly acute and represents the focus of this report.

EDs are responsible for over 140 million patient encounters in the US each year, as compared with approximately 1 billion outpatient clinic visits and 39 million inpatient stays.[4-6] Furthermore, EDs are the most common pathway for hospital admission in the United States. Thus, EDs are a critical interface between healthcare systems and the

communities they serve. Rapid diagnosis, risk stratification, and determination of the need for inpatient admission are core emergency medicine activities.[7] ED decisions have far-reaching consequences for patient morbidity and mortality, as well as healthcare costs.[8]

Data science and machine learning have the potential to augment clinician cognition in the ED by synthesizing vast quantities of clinical data available in the EHR and cross-referencing with exponentially increasing medical literature to identify subgroups of patients amenable to new, precision treatment.[9] However, significant technical and systemic barriers exist to allow collating, aggregating, and analyzing data in a meaningful and actionable manner. Furthermore, algorithms solely designed to detect certain biologic phenomena can become idiosyncratic reflections of what tests doctors tend to order. Algorithms trained purely on datasets have the potential to encode racial and gender biases, resulting in automations or magnification of such problems.[10] Understanding data surrounding these encounters is a tremendous opportunity to better characterize acute diseases, health care utilization, and ultimately public health.

In September of 2017, the National Institutes of Health (NIH) released a Request for Information (RFI) regarding data science research priorities (NOT-LM-17-006). A joint committee consisting of members of both the Society for Academic Emergency Medicine and American College of Emergency Physicians Research Committees, as well as selected research and health policy experts, were assembled to respond to this RFI and highlight priorities for data science research of relevance to emergency medicine. Content experts were recruited based on leadership positions in academic societies and clinical trial networks with current or a strong history of NIH research funding, prior publications or funding in project leveraging “big data” in emergency medicine applications, and/or significant publications and leadership in the area of emergency medicine health policy. If initial experts were not able to contribute, recommendations for their replacements were considered. Ultimately the group consisted of 12 contributors from 12 unique institutions geographically spread across the United States. The group was gathered rapidly in an *ad hoc* basis due to a short time frame from release of the RFI to the end of the comment period. As such, recommendations were developed via group email roundtable discussion rather than a modified Delphi approach, with all authors

contributing and agreeing on final recommendations. Priority areas focus on themes of fragmentation, access, fidelity, and formatting. The goal of this report is to disseminate research and policy targets identified by this group that, if properly addressed, will help overcome identified barriers and move big data science from "promise" to "practice." In June 2018, the NIH released their strategic plan (https://grants.nih.gov/grants/rfi/NIH-Strategic-Plan-for-Data-Science.pdf) which incorporated a number of our committees recommendations.

High priority areas for research and policy related to data science in the emergency department

1. *Develop improved methods for cross-platform identification and linkage of patients.* The need to ease cross-platform communication was identified by the group as both a clinical research priority and a critical clinical policy issue (which in turn has implications for observational, epidemiologic, and population research). Emergency, unscheduled patient care encounters involve multiple health care records and the records generated often lack interoperability, leading to significant challenges in transitions of care.[10,11] For example, a patient can easily generate 3-5 unique and unlinked medical records during a single emergency healthcare encounter (**Figure 1)**. A patient often presents to an independent outpatient setting using one electronic medical record system, is transported via one of several emergency medical services each using its own unique electronic charting system to an ED where the same patient may generate a third unlinked electronic chart, and ultimately is admitted to a hospital that may employ yet another EHR product. Downstream effects of such care become even more opaque when the patient is transferred from one ED or hospital to another, or to a post-discharge care setting (e.g., rehabilitation, nursing facilities). Such fragmentation of the medical record is the norm rather than the exception for most emergency care encounters, and inability to access data from multiple settings has the potential to systematically bias research findings though the introduction of selection bias based on how patients are identified and tracked longitudinally or measurement and verification bias based on clinicians' use of testing. Ultimately, erroneous application of big data techniques has the potential to adversely affect care. For example, if only data from a single, non-linked source is used,

filtering electronic health records for "complete data" in certain fields can introduce significant bias compared to claims databases which provide a more holistic view of longitudinal patient care.[12] If such cross-platform data are collected at all, it relies on labor-intensive manual chart abstraction or probabilistic linkage of records from multiple sources,[13] which can also introduce selection bias that is difficult to identify.[14,15] Future work in data science should identify scalable solutions to reduce fragmentation and promote access to data between systems or data aggregation across platforms, as well as ways for clinicians to easily view this information. Prescription Drug Monitoring Programs (PDMPs) represent one narrow example of how such systems may work. Future, broader programs would developing, deploying and adopting standards for interoperability, secure and private keys shared between medical records to allow unique linkage (such as an encrypted Globally Unique Identifier – GUID). Voluntary or mandated use of health information exchanges (HIEs) to create virtual complete records with adequate consideration of privacy protections[16,17] represents a laudable goal in this regard, but requires investment. Issues to date that have limited HIEs in their ability to fill this gap include incomplete community penetrance, leading to bias patient samples. In order to maximize their efficacy, federally mandated participation would be needed. Alternatively, a novel, unified, federal HIE could be developed, but would require significantly more investment. Finally, there are largely unexplored opportunities in combining standard medical care with nontraditional sources of data such as environmental exposures, social determinants of health, or patient consumer activity. However, interoperability of data collection systems will be paramount for these types of efforts to be conceivable.

2. *Create an NIH managed and maintained central, de-identified, open access databases for research purposes.* With increasing data collection, there is significant need for facile methods to seamlessly load increasing granular, de-identified, patient-level data into open access systems for the scientific community. The skeleton of such systems already exist through the Healthcare Cost and Utilization Project (HCUP) and various Centers for Disease Control databases, but limited data fields collected limit the hypotheses that can be tested using these resources. Privacy, ethical, and legal challenges need to be

surmounted. While the NIH has required public reporting of data for several years,[18] there has no single interoperable repository, no mechanism to do this easily, and no way to track when it is completed. The framework for such an approach exists in the NIH-supported genomics, proteomics, and metabolomics central repositories. However, sharing clinical data and linking disparate sources would build capacity to study complex disease states, long-term outcomes, and rare diseases that cannot be adequately studied with current methodology. Open access datasets also allow for improved *reliability of research* as well as external verification of statistical analyses. Recently there has been an increased concern regarding reproducibility in research evolving from genetic and microarray analyses.[19,20] The issues at play are complex, but range from vague methods, poor quality control, inconsistency in data reporting, and lack of statistical clarity which all culminate in an inability to reproduce research findings. Such "big data" problems are likely to affect clinical research as the deluge of information continues. Open access databases would allow for external validation of study findings using similar or orthogonal data analysis methods. Prior to creation of such repositories, however, adequate framework must be developed for their proper use and maintenance. Unless such systems remain facile and adequately supported, increased unfunded requirements of investigators (such as public reporting, ensuring data quality, and responsibility for response to queries) may inadvertently threaten data integrity, public perception of research reliability, and quell future research endeavors by investigators and patients alike. As evidence from across the information technology spectrum continues to demonstrate, data breaches seem to be a near-inevitability for purely online data repositories. Expansive datasets may be better maintained on isolated mainframes, with lock-and-key approval for access, under a model similar to the Healthcare Cost and Utilization Project (HCUP) database. We believe the NIH is the best poised to develop, operate, and maintain such as database to ensure a high quality, high fidelity, and secure data resource.

3. *Improve methodologies for visualization and analysis of intensively sampled data.* Increasingly, biometric data are accumulated by machines and recorded in an automated fashion. This can generate long streams of intensively sampled longitudinal

data. Examples include long electrocardiographic recordings, as well as sequential blood pressure, heart rate, or hemodynamic or biometric measurements. Patients often arrive with self-monitored data (e.g. heart rate from fitness monitors) as well. Standard methods and formats are needed to aggregate, synchronize, and annotate these time-varying data from multiple platforms. Methods to move, visualize, and analyze these data (particularly longitudinally) are also not well established. Future work should examine data management and analysis for such intensive longitudinal data. There also should be exploration of the meaning, significance and reliability of patient self-monitoring data for making treatment decisions. Individuals have already begun hacking and modifying their own devices, particularly glucose monitoring devices, demonstrating a field in which the medical community, for a variety of reasons is failing to meet patients' needs.[21]

Furthermore, akin to The Human Genome Project, there exists significant opportunity to create a *human imaging project* that includes linked phenotypic and anonymized imaging data that could be explored by researchers across the globe, enabling novel discovery from already acquired resources with due consideration of privacy and ethical issues. Finally, as technology continues to develop, files of huge size are being generated. We expect that as the number of types and intensity of sampling of these data increase, new compression techniques may be required for data transfer and/or storage. This may become particularly acute in the case of aggregate storage of longitudinal data of large numbers of patients, illustrating the need to partner with technology experts to develop not only strategic approaches but also technical solutions.[22]

4. *Develop methods to identify and standardize electronic medical record data quality.* Improved access to clinical and administrative information offers substantial opportunities for data science researchers. However, limited accuracy and reliability of such sources, especially those created in the emergency setting, may impair or misdirect such investigations.[17] Documentation that includes templates, copied text,[23] and automated advisories can lead to systemic misrepresentation and inconsistencies in medical records and administrative datasets. In the ED setting, rapidly evolving situations and a high flux of changing preliminary information contribute to

inconsistently accurate records. Improving the ability to ensure the fidelity of clinical datasets is an area ripe for investigation with limited attention to date. Future work in data science and medical informatics should include improved methods to detect and reduce problems with data quality in large datasets, and establishment of much needed standards. Examples of ensuring fidelity include back-end identification of data patterns indicative of potential systematic error, such as those that are repetitive, overly consistent, or anomalous in appearance. Front-end solutions to reduce error at the time of data creation are also desirable such as prevention of entering illogical or incompatible information.[24] Examples might include automated prompting for clinical verification of a positive pregnancy test result in a biological male patient (which could occur in the setting of testicular cancer) or a normal mental status in a patient who is intubated (which may occur immediately prior to extubation). Such examples demonstrate the need for broad stakeholder input and the development of improved human-computer interfaces, with a commitment to record integrity. Without the development of improved methods at the point of data entry, scientists are likely to have poor research quality data that is prone to erroneous findings and irreproducible or systemically biased studies. Additional consideration should be given to creation of "research-ready" documentation functionality within EHRs to ensure critical elements are routinely collected in a structured data format. Such an approach would be particularly valuable for accreditation or certification programs, which rely on clear demonstration of process measures (e.g., door to electrocardiogram time, or use of order sets for a given condition), and quality reporting, which require delineated numerators and denominators (e.g., proportion of low-risk chest pain patients who undergo stress testing) to derive accurate outcome data. By improving the upfront collection of information in the form of structured data, accuracy will be improved and the burden for back-end work will be diminished. However, implementing this will require a willingness of EHR vendors to deviate from the status quo – something that they have heretofore not displayed.

5. *Improve and utilize natural language processing for the more robust study of patient, provider, and systems level challenges.* Most data for ED encounters are contained in the history, physical exam, evaluation/management services, and imaging report components

of chart documentation. Unfortunately, these data are rarely structured in current medical records, and data are most often entered as free text or dictated text.[25] This creates a barrier to large-scale exploration of electronic medical records. It is highly likely these data elements are more reliable or more relevant to patient care given the *de facto* emphasis placed by the clinician on communicating thought process through the use of free text. Absent the ability to implement upfront utilization of structured data at intake, better methods to work with unstructured data and seamlessly convert it to a usable format are needed. While there are a number of technological solutions have expanded the potential to achieve this, such an approach has yet to be integrated into the clinical arena for routine data management.[3,26-27] Future work should examine how to structure abundant free-text data from encounters into analyzable forms to preserve the richness of these data as opposed to forcing artifactual discrete data field entry. Novel methods including two-step "smart" processing should also be explored, whereby discrete data points that correspond to a diagnosis or criteria for study inclusion are automatically transformed (e.g., echocardiogram report of an ejection fraction = 35% is converted to a diagnosis of heart failure with reduced ejection fraction, or a potassium of 6.5 mmol/L is interpreted by the processor as hyperkalemia).

6. *Develop and utilize syndrome or complaint-based taxonomies of disease.* Patient encounters in emergency medicine are poorly characterized using common taxonomies for disease.[28,29] As an example, a patient may be classified by a final diagnosis mapped to an ICD-10 code (e.g., gastroesophageal reflux). However, this code does not reflect the initial symptoms or physiological syndrome that led to an ED visit (e.g., chest pain). Use of ICD-10 or other diagnostic coding mechanisms is therefore a poor manner to assess whether utilization or testing was appropriate (e.g., stress testing or CT scan) for a given presentation. Syndromic taxonomies have been developed by the Centers for Disease Control and the National Library of medicine, including SNOMED-CT, [30]that could be used as the basis for such taxonomies. Future work in the data sciences should develop standards for how research findings based on post-hoc diagnoses made after diagnostic testing and workup compare to an undifferentiated patient population. For instance, studies of patients with an ICD-10 diagnosis of sepsis could compare their

results to an unselected cohort of patients meeting consensus criteria for sepsis in the ED or who present with a vague complaint (e.g., fever, or body aches) that may or may not ultimately be coded as sepsis. Aforementioned improvements in natural language processing focused on chief complaint may be particularly useful in this regard. This would enable a better understanding of the diagnostic decision making at the provider level, and help to interpret the accuracy of relatively non-specific criteria that can, by virtue of being tied to performance metrics, trigger unnecessary or even inappropriate care (e.g., administering large amounts of fluid to a patient solely based on sepsis criteria to avoid a perceived or actual penalty). The EM Common Core Model may also serve as a framework for such a system.

7. *Develop a practical and ethical framework to leverage electronic systems for data collection in controlled trials.* Data science can often use naturally occurring variability to infer differences between groups of interventions, but such findings are limited by unrecognized confounding. Furthermore, systematic selection bias can easily be introduced and may not be adequately evaluated using typical methods to address missingness.[31] More reliable and accurate confirmation of therapeutic effects within large, ongoing clinical and administrative datasets may require incorporating patient-level or site-clustered allocation to an intervention by random or quasi-random methods. Current methods of integration require labor-intensive human-level data abstraction and serve as an impediment to the seamless conduct of clinical trials. This impediment is particularly important outside of academic medical centers and potentially contributes to systemic bias of research findings. Implementation of such integration would enhance data capture, offering novel methods to increase protocol fidelity (e.g., pop-ups or text messages to patients to document pain levels or to nursing to chart updated vital signs) and perhaps expand the type of effects that can be assessed. For example, by incorporating patient flow data into heath records, we may be able to automate modeling of patient care efficiency (e.g., throughput times for service, waits in queue), and other data relevant to operational improvements in the ED, while providing a readily accessing test environment to study alternative approaches to care delivery (e.g., fast track, team triage, etc).

8. *Explore technologies to help enable conduction of clinical trial in the emergency setting by streamlining subject identification*. Achieving the aforementioned defragmentation of electronic medical records can lead to improved methods to identify eligible subjects for clinical trials in the emergency care setting. Examples of these that could be explored include a national database of pre-encounter study consent (i.e., patients consent in advance to participate in an emergency care trial where they may be unable to consent at the time of their acute disease), use of videotaped presentations to provide information necessary for informed consent, matching of patients to potential studies via background electronic medical record analysis, and automated notification of patients and providers about eligibility for studies.[33-36] The use of registry-based randomized controlled trials,[37] which leverage preexisting registries (which are relative low cost and internally valid) to identify patients or institutions for randomization, may also be an option. As high-quality data are already being collected, such trials decrease the need for data collection and therefore cost, and should be part of the pragmatic data science toolbox for learning health systems of the future. However, at present, such methods remain underdeveloped and inconsistently applied. Funding for pilot studies could optimize procedures, leveraging the strengths of novel data science applications.

9. *Train emergency care clinicians in data science and data scientists in emergency care medicine*. Development and funding of targeted K-level training grants beyond the K01 mechanism will be required to develop researchers, especially clinician-scientists, in data science. This data scientist workforce must be capable of both creating data science methodology and applying data science approaches to the emergency care setting. Application of data science approaches should include both traditional bioinformatics and clinical health informatics, as well as public health informatics. Programs should also be developed to train a new type of clinician in data science. For success, clinicians must be intimately involved in curating data, choosing outcomes to predict, and ultimately building and rigorously testing algorithms to ensure data science research remains firmly rooted in the realities of clinical care. To help prepare for this, such training ideally would begin at the undergraduate level and continue into graduate education in medical

schools as well as computer science and engineering pre- and post-doctoral programs. However, to operationalize in a meaningful way, dedicated fellowship training and on-going faculty career development programs would be needed.

Conclusion

The exponential growth of healthcare data carries enormous promise for the better understanding of health and disease, with the potential for tangible benefits for patients and providers alike. Such advances are in danger of being stalled by a number of theoretical and practical barriers. Coordinated research and policy approaches may help lower some of these barriers to help fulfill the promise of big data in emergency care.

Table 1: Summary of high priority research and policy recommendations, with examples of solutions and potential pitfalls to implementation or adoption

Figure 1: Example of how a single emergency care encounter can generate multiple unique, unlinked healthcare records juxtaposing current and proposed paradigms of data collection (PCP – Primary care provider; ED – Emergency Department; EMS – Emergency Medical Services; LTAC – Long term acute care; CMS – Center for Medicare and Medicaid Services)

Literature Cited

(1) Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst.* 2014;2:3.

(2) Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption Factors of the Electronic Health Record: A Systematic Review. *JMIR Med Inform.* 2016;4:e19.

(3) Janke AT, Overbeek DL, Kocher KE, Levy PD. Exploring the Potential of Predictive Analytics and Big Data in Emergency Care. *Ann Emerg Med.* 2016;67:227-236.

(4) https://www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf. Accessed June 15, 2018.

(5) https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2014_namcs_web_tables.pdf. Accessed June 15, 2018.

(6) https://www.hcup-us.ahrq.gov/reports/statbriefs/sb235-Inpatient-Stays-Age-Payer-Trends.pdf. Accessed June 15, 2018.

(7) Pines JM, Lotrecchiano GR, Zocchi MS et al. A Conceptual Model for Episodes of Acute, Unscheduled Care. *Ann Emerg Med.* 2016;68:484-491.

(8) Galarraga JE, Pines JM. Costs of ED episodes of care in the United States. *Am J Emerg Med.* 2016;34:357-365.

(9) Obermeyer Z, Lee TH. Lost in Thought - The Limits of the Human Mind and the Future of Medicine. *N Engl J Med.* 2017;377:1209-1211.

(10) McMurray J, Hicks E, Johnson H, Elliott J, Byrne K, Stolee P. 'Trying to find information is like hating yourself every day': the collision of electronic information systems in transition with patients in transition. *Health Informatics J.* 2013;19:218-232.

(11) Emergency Department Transitions of Care - A Quality Measurement Framework Final Report . *National Quality Forum.* 2017.

(12) Weber GM, Adams WG, Bernstam EV et al. Biases introduced by filtering electronic health records for patients with "complete data". *J Am Med Inform Assoc.* 2017;24:1134-1141.

(13) Newgard CD, Zive D, Jui J, Weathers C, Daya M. Electronic versus manual data processing: evaluating the use of electronic health records in out-of-hospital clinical research. *Acad Emerg Med.* 2012;19:217-227.

(14) Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. *BMC Med Res Methodol.* 2014;14:36.

(15) Harron KL, Doidge JC, Knight HE et al. A guide to evaluating linkage quality for the analysis of linked data. *Int J Epidemiol.* 2017;46:1699-1710.

(16) Kum HC, Ahalt S. Privacy-by-Design: Understanding Data Access Models for Secondary Data. *AMIA Jt Summits Transl Sci Proc.* 2013;2013:126-130.

(17) Balas EA, Vernon M, Magrabi F, Gordon LT, Sexton J. Big Data Clinical Research: Validity, Ethics, and Regulation. *Stud Health Technol Inform.* 2015;216:448-452.

(18) Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov - The Final Rule. *N Engl J Med.* 2016;375:1998-2004.

(19) Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet.* 2006;7:55-65.

(20) Bustin SA, Huggett JF. Reproducibility of biomedical research - The importance of editorial vigilance. *Biomol Detect Quantif.* 2017;11:1-3.

(21) Mann, M. “This Diabetes Activitist Hacked Her Medical Device and Made an Artificial Pancreas.” https://motherboard.vice.com/en_us/article/aekkyj/this-

diabetes-activist-hacked-her-medical-device-and-made-an-artificial-pancreas. Accessed June 15, 2018.

(22) Francescon R, Hooshmand M, Gadaleta M, Grisan E, Yoon SK, Rossi M. Toward lightweight biometric signal processing for wearable devices. *Conf Proc IEEE Eng Med Biol Soc.* 2015;2015:4190-4193.

(23) Patterson ES, Sillars DM, Staggers N et al. Safe Practice Recommendations for the Use of Copy-Forward with Nursing Flow Sheets in Hospital Settings. *Jt Comm J Qual Patient Saf.* 2017;43:375-385.

(24) Daymont C, Ross ME, Russell LA, Fiks AG, Wasserman RC, Grundmeier RW. Automated identification of implausible values in growth data from pediatric electronic health records. *J Am Med Inform Assoc.* 2017;24:1080-1087.

(25) Kimia AA, Savova G, Landschaft A, Harper MB. An Introduction to Natural Language Processing: How You Can Get More From Those Electronic Notes You Are Generating. *Pediatr Emerg Care.* 2015;31:536-541.

(26) Polnaszek B, Gilmore-Bykovskyi A, Hovanes M, Roiland R, Ferguson P, Brown R, Kind AJ. Overcoming the Challenges of Unstructured Data in Multi-site, Electronic Medical Record-based Abstraction. Med Care. 2016 Oct; 54(10): e65–e72.

(27) Luo L[1], Li L[1], Hu J[1], Wang X[1], Hou B[1], Zhang T[1], Zhao LP[2]. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system.BMC Med Inform Decis Mak. 2016. 16:114.

(28) Griffey RT, Pines JM, Farley HL et al. Chief complaint-based performance measures: a new focus for acute care quality measurement. *Ann Emerg Med.* 2015;65:387-395.

(29) Berdahl C, Schuur JD, Fisher NL, Burstin H, Pines JM. Policy Measures and Reimbursement for Emergency Medical Imaging in the Era of Payment Reform: Proceedings From a Panel Discussion of the 2015 Academic Emergency Medicine Consensus Conference. *Acad Emerg Med.* 2015;22:1393-1399.

(30) Lee D, de Keizer N, Lau F, Cornet R. Literature review of SNOMED CT use. *J Am Med Inform Assoc.* 2014:21(e1):e11-9.

(31) Haneuse S, Daniels M. A General Framework for Considering Selection Bias in EHR-Based Studies: What Data Are Observed and Why? *EGEMS (Wash DC).* 2016;4:1203.

(32) Saver JL, Starkman S, Eckstein M et al. Methodology of the Field Administration of Stroke Therapy - Magnesium (FAST-MAG) phase 3 trial: Part 2 - prehospital study methods. *Int J Stroke.* 2014;9:220-225.

(33) Furyk J, McBain-Rigg K, Watt K et al. Qualitative evaluation of a deferred consent process in paediatric emergency research: a PREDICT study. *BMJ Open.* 2017;7:e018562.

(34) O'Malley GF, Giraldo P, Deitch K et al. A Novel Emergency Department-based Community Notification Method for Clinical Research Without Consent. *Acad Emerg Med.* 2017;24:721-731.

(35) Offerman SR, Nishijima DK, Ballard DW, Chetipally UK, Vinson DR, Holmes JF. The use of delayed telephone informed consent for observational emergency medicine research is ethical and effective. *Acad Emerg Med.* 2013;20:403-407.

(36) Brienza AM, Sylvester R, Ryan CM et al. Success Rates for Notification of Enrollment in Exception From Informed Consent Clinical Trials. *Acad Emerg Med.* 2016;23:772-775.

(37) Mathes T, Buehn S, Prengel P, Pieper D. Registry-based randomized controlled trials merged the strength of randomized controlled trails and observational studies and give rise to more pragmatic trials. *J Clin Epidemiol.* 2017.

Recommendation	***Examples***	***Pitfalls***
1. Develop improved methods for cross-platform identification and linkage of patients	Global unique identifier (GUID)	Security / Privacy
2. Create central, de-identified, open access databases	NIH –omics repositories	Unfunded mandates, system maintenance
3. Improve methodologies for visualization and analysis of intensively sampled data	Continuous telemetry, fitness trackers	File size, data storage, proprietary restrictions
4. Develop methods to identify and standardize electronic medical record data quality	Identification of template overuse, prevention of illogical data entry	Evolving, unreliable history
5. Improve and utilize natural language processing	Leverage richness of natural language over discrete data fields	Clinician level and regional variations
6. Develop and utilize syndrome or complaint-based based taxonomies of disease	Chest pain rather than gastroesophageal reflux disease	Billing tied to diagnosis codes
7. Develop a practical and ethical framework to leverage electronic systems for controlled trials	Patient level or site clustered randomization	Overreliance on statistical modeling and inference
8. Explore technologies to help enable clinical trials in the emergency setting	National database of pre-encounter consent	Practical framework, time-sensitivity
9. Train emergency care clinicians in data science and data scientists in emergency care medicine	K08, K23, and K24 mechanisms	Dissociation of clinical practicalities from data analysis