

Clinical data mining and research in the allergy office

Dan Dalan

Chair Health Informatics, Technology & Education
Committee American Academy of Allergy, Asthma &
Immunology, Fargo, North Dakota, USA

Correspondence to Dan Dalan, MD, Allergy and
Asthma Care Center 4776, 28th Avenue South #201,
Fargo, ND 58104, USA
Tel: +1 701 232 8788;
e-mail: dandalan@medicine.nodak.edu

**Current Opinion in Allergy and Clinical
Immunology** 2010, 10:000–000

Purpose of review

More data are anticipated from the expected increase in use of electronic health records (EHRs). Upcoming initiatives require reporting of quality measures, meaningful use of clinical decision support, alert systems, and pharmacovigilance – knowledge resulting through use of EHRs. Data mining is a new tool that will help us manage information and derive knowledge from these data, and is a part of evolving new disciplines of informatics and knowledge management.

Recent findings

Studies are reported from smaller clinic data marts to larger repositories and warehouses in various health systems, biomedical registries, and the medical literature on the Internet. Data mining technologies show promise and challenges. Outcome measures as structured data and narrative text can be mined with human assistance and newer automated natural language processing software. Despite advances, the growing diversity of clinic EHRs lack integration and interoperability with Internet-based biomedical databases.

Summary

Allergists have the capability to mine clinic EHRs to discover new information, which may be hidden in charts. A central allergy computer can serve not just as a registry but also allows functionalities to enable EHRs' meaningful use. Harmonization of technological and organizational standards will allow seamless use of new natural language processing (NLP) tools and ontologies through a semantic web.

Keywords

allergy clinic research, data mining, informatics, knowledge management, semantic web

Curr Opin Allergy Clin Immunol 10:000–000
© 2010 Wolters Kluwer Health | Lippincott Williams & Wilkins
1528-4050

Introduction

We will accumulate more data in our clinics from the increase in use of EHRs and explosive growth of medical databases [1,2,3^{••}]. Although computerization of our daily clinic workflow will improve efficiency, it does not guarantee meaningful use. Additional effort in the form of knowledge acquisition and use of informatics to improve patient care is recommended by the Institute of Medicine [4]. Traditional manual analysis of our current charts now will need to be coupled with methods of computer analysis such as data mining [2], which can help us cope and manage the additional data, information, and can help create new knowledge [5[•]]. With this comes the prospect of additional improvement of both clinics and patient care outcomes.

Much is written about clinical data mining between larger universities, health systems, and interdisciplinary biomedical databases, but few from the allergic clinic. A main focus of research and review articles in this field asks for the integration and facilitation of data mining, and

emphasizes correlation of phenotypic data from clinics, environmental, genomic, and other omic data warehouses now becoming more available. Next generation tools available include data mining, implementation of knowledge management systems, and related NLP, which efficiently gather new knowledge from these interactions from several disciplines [2,3^{••},4]. Many [6^{••},7[•],8^{••}] discuss the difficulty with integration of interoperable databases from disparate groups, a main hurdle for which much progress is being made through organizational changes, refinement of automated computer machine learning of tasks in NLP, and use of a semantic web [8^{••},9]. Despite some progress, they are not in common use [5[•]].

My review will focus primarily on allergy office-based activities and research that may result in practical use or report new findings. Although small clinic data mart studies do not reflect the population, these examples will help introduce to allergists a new discipline. I hope to stimulate reporting of similar findings, a required validation process, and promote the formation of a larger

allergy data warehouse for research of more robust data to garner knowledge reflecting more of the population.

Electronic health records (EHRs) technological standards are becoming more interoperable due to requirements [10[•]]. Recommendations for integration and interoperability of these systems are suggested through a standard infrastructure and use of relational databases [6^{••}]. Most EHRs now are capable of information retrieval through report functions. These are prerequisite activities prior to data mining, which enable us to derive new knowledge hidden in our own charts. However, computerized data mining can be a more challenging task in EHRs and online databases. Software for NLP required to do this is complicated and difficult to use, and is mostly available in the institutions that created them [11[•]].

I will review recent examples of data mining using NLP techniques in other medical settings, and then introduce the semantic web. I will end with a look at the potential impact of advancements in clinic information technology in the future of the allergy clinic.

Definitions

Data mining is the analysis of data for relationships that have not previously been discovered and is becoming an increasingly important tool to transform data into information, then knowledge [10[•]]. Data are raw facts stored as characters, symbols, or words when unprocessed, and are not useful in our clinic practice. For example, the numbers 4, 9, 3, 0, 1 are just data points, but ICD-9-CM of 493.01 codes for exacerbation of extrinsic asthma. This is new information obtained by retrieval and processing data in EHRs. Additional analysis and interpretation by the allergist of the narrative text may explain what caused the exacerbation, and this information needs to be extracted for the data to become information, then knowledge. A new consult visit is data mining when we conclude the diagnosis and treatment suggestions in our report.

Lavindrassana *et al.* [3^{••}] summarized the literature description of data mining into strategies of human supervised data mining, akin to hypothesis testing, unsupervised exploratory data analysis, or a combination of the two.

After mining, knowledge management means getting the right information to the right people at the right time utilizing people, processes, and technology to help make better decisions for the organization [12[•]], allowing them for meaningful use. We send our consultation evaluation and treatment to our referring physicians, and this is knowledge management, when done electronically and

shared at the point of care, such as through Health Level 7's (HL7) Clinical Document Architecture [13[•]], soon to be a requirement [10[•]] for administrative patient care, report, and sharing of our findings online to interested parties.

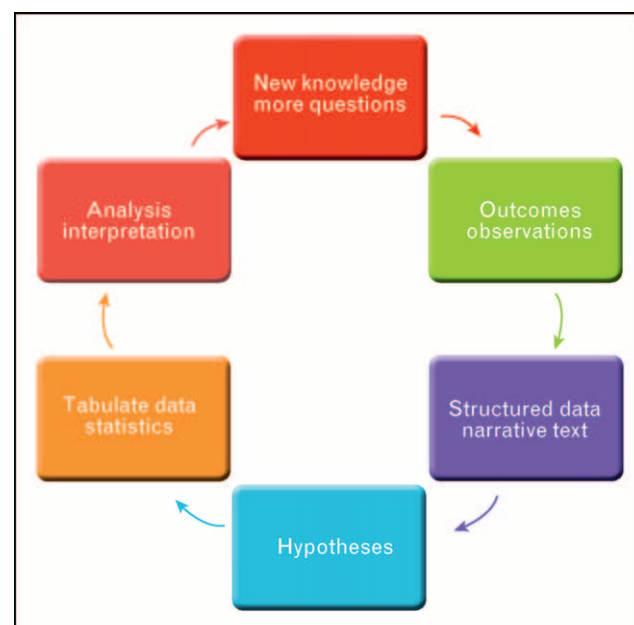
Informatics is the science of computerized information management. It involves the collection, analysis, storage, and integration of data as well as the dissemination, presentation, and safeguarding of information [14]. There is an amalgamation of technology, knowledge, hardware, software, and people ware [14]. Informatics describes data warehousing and mining, [15]. A typical clinic data mining research schema is displayed in Fig. 1 [16^{••}].

Natural language processing software attempts to analyze and retrieve data from EHRs without human intervention. It considers the sentence structure, syntax, meaning, and context to accurately process, and then extract information from the narrative text. Several advancements suggest a hybrid automated and human analysis [7[•], 17^{••}], and allow more accurate processing.

Ontology is set-of-concept-definitions for the purpose of enabling knowledge sharing and reuses [18].

Hayrinen *et al.* [19[•]] define EHRs as repositories of patient data, stored and exchanged securely, accessible,

Figure 1 Representation of informatics research in an allergy clinic



Data mining of clinic EHRs chart through hypothesis testing and exploratory data analysis. EHRs query function can create reports of outcome measures, which you tabulate, then statistically analyze [16^{••}] and discover new knowledge. EHRs, electronic health records.

the primary purpose of which is to support quality and efficient care. Current research on EHRs use shows that most documentation of outcomes measure is by unstructured narrative texts and not easily usable by computer programs [20[•]]. However, structured texts are more easily processed, examples of which are numerical values of outcome measures or tests. The challenge with narrative texts lies in the varying context and semantics, which are difficult for searching, summarization, and statistical analysis. Structured text and data can be easily coded, but narrative texts require additional processing through NLP, in which the textual information is extracted, linked to standard terminologies, and then coded, which facilitates analysis by computers.

The semantic web is a common online framework that allows data to be shared and reused across application, enterprise, and community boundaries, in which the meaning or semantics of information and services is defined, making it possible for the web to comprehend and satisfy the requests of people and machines to use the content on the Internet [8^{••},9].

Manual computer data mining in the allergy clinic

Because automated data mining with NLP is not widely available in clinics, we analyze mostly structured data in our EHRs. We can data mine our study or EHRs' chart by tabulating data suggested by Marshall and Bahna [16^{••}] and proceed with the research schema seen on Fig. 1. EHRs can retrieve information of mainly any coded numerical test values or structured data in our charts, and produce a spreadsheet for analysis through their query or report function. Many of us share similar observations in our allergy clinics.

I noticed increased skin prick test (SPT) reactions to *Cephalosporium acromonium* mold in our panel, which led to research and hypothesis testing [21]. Incidence of this mold reaction over time revealed a Gaussian curve coincidental with exposure and sensitization after a major flood. We used global positioning satellite (GPS) technology with patient demographics to analyze and then discover that these mold allergic patients were at a higher risk of sensitization when they are within the flood plains (Fig. 2). These findings are important in allergy health-care decisions in such flood prone geographic and/or farming areas.

The rhinitis outcomes questionnaire (ROQ) is a validated survey recommended for use in health outcomes [22]. Once processed, they become coded structured numerical scores. I tabulated values over time of patients' ROQ and showed statistically significant improvement after the first, and each of the subsequent visits over 5 years.

This and similar surveys provide patient centric results recommended for quality measures reporting [3^{••},10[•]].

Correlation of the National Allergy Bureau (NAB; www.aaaai.org/NAB) pollen counts to pollen allergic patients' ROQ – clinic phenotype environment – is a common sense study that shows expected associations. Coupled with other demographic data and adherence measures [23], I was able to show an increase observance to treatment during the pollen season by e-mailing an intervention cohort. These can be alert and decision support functions in EHRs.

The allergist routinely analyzes and interprets charts, and with only a few additional steps, can retrieve and mine new knowledge that will help improve the evaluation and treatment of patients. With expected more EHRs' use come more opportunities with research, documentation of more diverse outcomes measures, and automation opportunities.

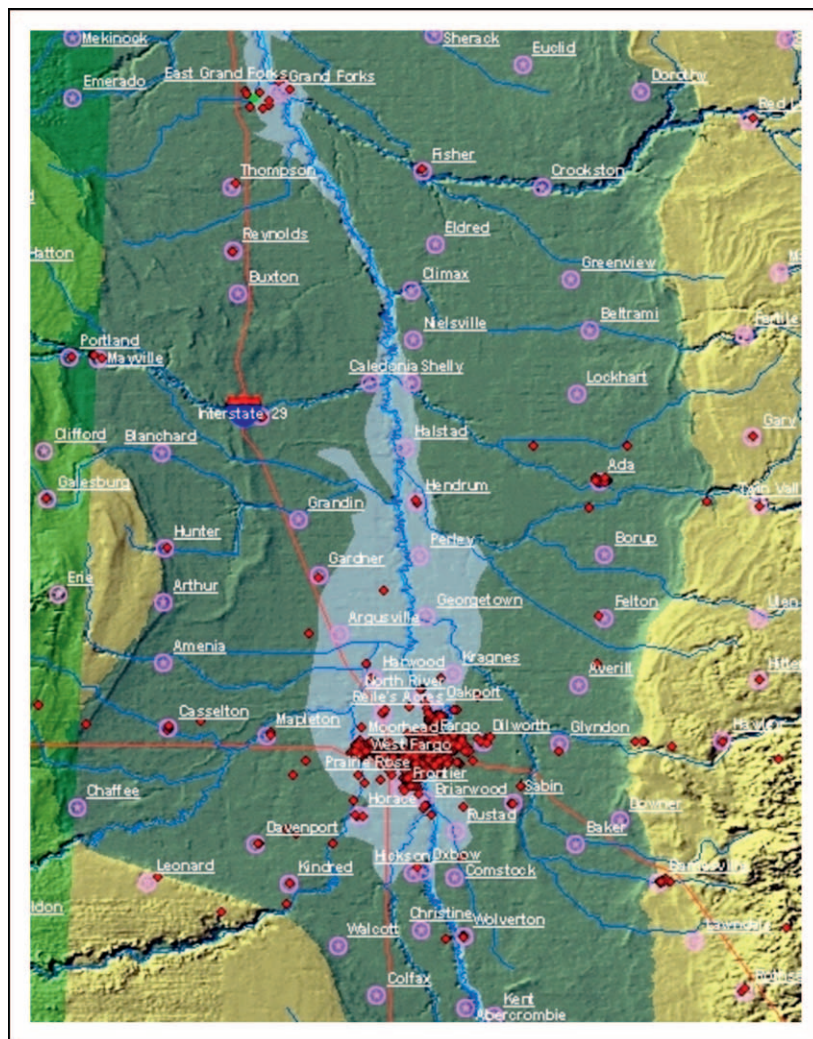
Automated data mining studies with NLP

Lavindrassana *et al.* [3^{••}] reviewed the literature and reported on the practice of automated clinical data mining. Most manuscripts' study purpose was for quality of care, data exploration, prognostic evaluation, and NLP. They, as well as most authors I reviewed, concluded that more efforts are needed to get acceptance of these new tools from healthcare professionals. Reproducible, easy to use, and understanding of software extraction of the information are needed. In most instances, human supervised procedures are best, a good reason why automated data mining with NLP is not in common use. Regardless, I chose selected examples on narrative text studies and how they may be used in our clinics. They actively search for clinic data and external factor correlations in new arenas of the Internet, the medical literature, and large data repositories.

Chen *et al.* [24] detected a significant correlation in trends and fluctuation of asthma medication use over 10 years in a healthcare system when compared with the medical literature, a classic phenotype literature study. As time proceeded, fewer articles appeared about theophylline, reflecting decreased use in the clinics and vice versa, with fluticasone–salmeterol combination. Seyfried *et al.* [25[•]] were able to shift through EHRs' narrative text for clinical trial patient factors that met selection criteria, a procedure helpful for clinics screening for such studies. These are practical activities we need and do now in our clinics, and as our volume of work and data grow, automation of these processes would be helpful.

Online concurrent monitoring is a large public health interest, with many diseases of importance. Conway *et al.*

Figure 2 Contour map of the flood plains in the Red River Valley revealed through use of global positioning satellite software



Outcomes measure is *Cephalosporium acremonium* mold sensitivity, a higher risk event if patient residence (red diamonds) is within the flood plains, colored blue. An example of clinic phenotype–external environment correlation.

[26[•]] were able to classify a disease outbreak report on BioCaster, a global health newscast that scans online information reports concerning infectious disease. Eysenbach [27^{••}] has pioneered real-time monitoring of patient symptoms when he first reported using *Google* Ad Words of influenza symptoms to predict the season. I tested this phenomenon by tracking patient allergy symptoms in the same *Google* tool (limited to a few miles within our counting station), and they correlated with the NAB ragweed pollen counts. *Google* hits for these symptoms essentially predicted the start of the ragweed pollen emergence, peak, and end of the season [28]. This is a field that is vast, potentially rich with data and information, which may correlate with our allergy patient's phenotype data contained in our EHRs. This correlation with the environment is always important in the allergic manifestation of disease, and as we are discovering new

sources of knowledge that may improve our evaluation and treatment methods.

In addition to and complimentary to phenotype–environment correlation studies are the correlations to genotypes (patient and epitope) and other omics data warehouses for allergy, now more and more reported. All studies here describe the existence of such resources and the possibilities of investigations. Schein *et al.* [29] provided an overview of databases on allergenic proteins and described state-of-the-art informatics tools to identify group allergens and detect areas that may account for common patterns of IgE binding and cross-reactivity. Samarghitean and Vihinen [30[•]] promote the use of many primary immunodeficiency databases now available online, facilitated by novel informatics tools. Gendel [31^{••}] described different available food allergen databases. However,

because of content nomenclature, organization, and database accessibility, he describes barriers for users that prevented data sharing and integration. He suggests a more semantic allergen nomenclature and use of a semantic web. Here lies the difficulty in the phenotype–environment–genotype process of investigating for many potential new and important findings.

Despite advances, methods using human supervision are still better than artificial intelligence techniques, but they are becoming useful adjuncts.

Future clinic health informatics technologies and the semantic web

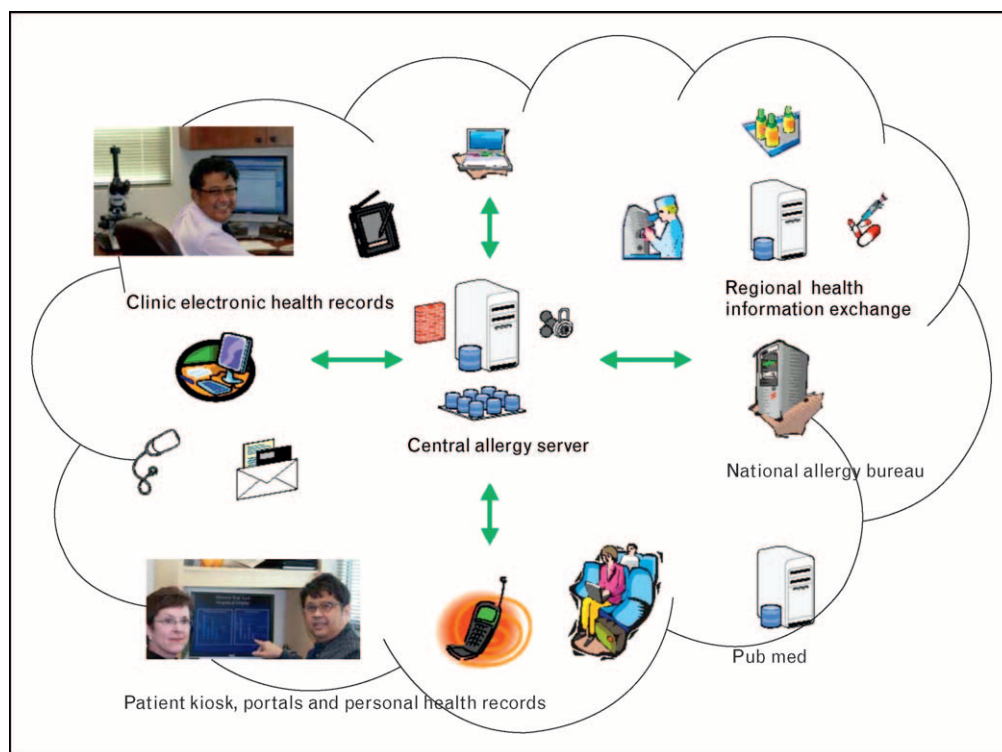
Sittig [32] describes the sporadic availability of technologies mentioned here, and many changes are coming that may include real-time clinical decision support systems, wireless monitoring devices, larger integrated, patient-level, phenotypic and genotypic databases with intelligent data mining capabilities. A schematic of these online tools from patients–clinics–data warehouses–healthcare systems is depicted in Fig. 3. I propose a central allergy registry as a proof-in-concept prototype. Seamless interactions depicted by double arrows here between personal patient devices, clinic computers, and central servers are major mechanisms of successful data mining.

Go Web is a next generation semantic search engine (www.gopubmed.org/goweb) [8^{**}]. It combines keyword-based web exploration with text-mining and ontologies to instruct large results obtained to a smaller more relevant set, facilitating answers of questions asked. They provided three benchmark studies to support their claim in gene associations and symptoms–protein–disease associations to demonstrate not only the power of the idea but also its limitations. So, the future is now here, although much of this new instrument has yet to be fully tested.

Points of care patient-centric approaches are facilitated with these informatics technologies. The allergist should lead the way on creating allergy-related organizational standards such as clinical decision support-alert systems and computerized physician order entry that will be required on EHRs [10^{*}]. We can then facilitate health information exchange through online systems and data warehouses, a current regional, national and international trend [3^{**}].

I reviewed the asthma IQ [33^{**}], a central repository of patient outcomes capable of quality measures reporting, but it has the same limited interoperability with clinic EHRs. Asthma IQ or an equivalent allergy registry can start to serve as a central warehouse for adding more data

Figure 3 Proof-in-concept prototype of a central allergy registry



An e-allergist using current and future online tools from patients–clinics–data warehouses–healthcare systems.

and functionalities, but the process will require changes in infrastructure and an organizational culture transformation to meet our future subspecialty needs in technology. Several organizational issues were discussed by Goble and Stevens [6**] and Gendel [31**] as hindrances to data mining of large data warehouses relating to integration, interoperability, and communications. We discussed similar hurdles but with solutions [34*] on organizational implementation of health informatics systems [35], as the continuum from personal health records, EHRs, and regional and world health information networks form.

Conclusion

We discussed several data mining examples in clinic and research as a way to discover new meaning in data that may be hidden in our charts. New informatics hi-tech tools may improve our delivery of healthcare and share information in a more efficient manner. A central allergy computer can serve not just as a data warehouse but also allows functionalities to enable EHRs' meaningful use. Harmonization of technological and organizational standards will allow seamless use of new NLP tools and ontologies through a semantic web. Research methods using human supervision are still better than artificial intelligence techniques of NLP, but they are becoming useful adjuncts. We can utilize what we have now and practice mining of our current clinic EHRs to be prepared for the future.

Acknowledgements

I would like to thank Joan Marie Dalan for her support in the preparation and editing of this manuscript, and all members of the electronic Joint Task Force (eJTF) on Standards of the Academy and College of Allergy, Asthma, and Immunology for the inspiration and dialogue in our committee on these matters.

I declare no conflicts of interests, financial or otherwise.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

Additional references related to this topic can also be found in the Current World Literature section in this issue (pp. 000–000).

- 1 American Recovery and Reinvestment Act of 2009. <http://www.hhs.gov/recovery/overview/index.html>. [Accessed December 2009]
- 2 Lavarac N. Selected techniques for data mining in medicine. *Artificial Intelligence in Medicine* 1999; 16:3–23.
- 3 Lavindrassana J, Cohen G, Depeursinge A, *et al.* Clinical data mining: a review. *Yearb Med Inform* 2009; 121–133.
The authors provide us with the definitions and the most current review about the topic and is a primary research analysis of current literature, the status-of-practice, and challenges ahead.
- 4 Institute of Medicine. <http://www.iom.edu>. [Accessed December 2009]
- 5 Karsten H, Suominen H. Mining of clinical and biomedical text and data: editorial of the special issue. *Int J Med Inform* 2009; 78:786–787.
This editorial discussed and focused this topic into perspective, several influential primary researches in this special issue that are discussed in this review.

- 6 Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008; 41:687–693.
The authors discuss the difficulty with integration of interoperable databases and the refinement of automated computer machine learning of tasks in NLP and use of a semantic web as potential solutions.
- 7 Roberts A, Gaizauskas R, Hepple M, *et al.* Mining clinical relationships from patient narratives. *BMC Bioinform* 2008; 9 (Suppl 11):1–17.
They report an additional refinement of NLP in the field of oncology using supervised data mining and technology emphasizing studies like this to support genotype-meets-phenotype informatics and move to translational research.
- 8 Dietze H, Schroeder M. Go Web: a semantic search engine for the life science web. *BMC Bioinform* 2009; 10 (Suppl 10):1–13.
The authors report successful use, with limitations, of a next generation semantic search engine backed by their pilot study data and free use of Go Web for noncommercial users at www.pubmed.org/goweb.
- 9 Berners-Lee T, Hendler J, Lassila O, *et al.* The semantic Web. *Sci Am* 2001; 284:34–43.
- 10 Centers for Medicare and Medicaid Services and the Office of the National Coordinator. <http://www.cms.hhs.gov/apps/media/press/release.asp?Counter=3561> ONIC. [Accessed January 2010]
This governmental agency issues regulations proposing a definition of meaningful use and setting standards for EHRs.
- 11 Jagannathan V, Mullett C, Arbogast J, *et al.* Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *Int J Med Inform* 2009; 78:284–291.
The authors describe limited use of even available NLP, and most are in institutions that produced them.
- 12 LaTour KM, Maki SE, editors. Health information management. Concepts, principles, and practice. 3rd ed. Chicago: American Health Information Management Association; 2010. pp. 111–159.
Informatics and data mining are not only defined, but practical current use and future applications in healthcare are also well presented.
- 13 Health level 7 clinical document architecture, CDA. <http://www.hl7.org/documentcenter/public/faq/cda.cfm>. [Accessed December 2009]
An organization that provides a framework for the exchange, integration, sharing, and retrieval of electronic health information that support clinical practice and the management, delivery, and evaluation of health services commonly used in the world.
- 14 Shortliffe E, Ciminino J. Biomedical informatics. Computer applications in healthcare and biomedicine. 3rd ed. Secaucus, New Jersey: Springer-Verlag; 2006. p. vii.
- 15 Wan T. Healthcare informatics research: from data to evidence-based management. *J Med Sys* 2006; 30:3–7.
- 16 Marshall G, Bahna S. Statistics for clinicians. *Ann Allergy Asthma Immunol* 2009; 103 (Suppl 1):2–3.
The author discussed and recommended data tabulation and analysis by use of available statistical software.
- 17 Wang X, Hripsak G, Friedman C. Characterizing environmental and phenotypic associations using information theory and electronic health records. *BMC Bioinform* 2009; 10 (Suppl 9):1–7.
This article used combined supervised hybrid exploratory data mining, and more refined NLP analysis, of discharge summaries looking for associations between environmental factors and disease phenotypes and found new correlations with symptoms, hypertension, and antihypertensive medicines important in pharmacovigilance.
- 18 Gruber T. A translational approach to portable ontology specifications. *Knowledge Acquisition* 1993; 5:199–220.
- 19 Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008; 77:291–304.
The authors provide an up-to-date definition and how EHRs are used in this global review of the literature.
- 20 Meystre S, Savova G, Kipper Schuler K, *et al.* Extracting information from textual documents in the electronic health record: a review of recent research. *IMIA Yearbook Med Inform* 2008; 47 (Suppl 1):128–144.
This is a review and primary research of the literature on the type of outcomes measures contained on EHRs, which is mostly narrative text, as well as the challenges in extracting data through the use of NLP.
- 21 Dalan D, Brooks W, Levetin E. Effects of climate and flooding on mold & pollen sensitization [abstract]. *J Allergy Clin Immunol* 2007; 119.
- 22 Dalan D. Rhinitis outcomes questionnaire in allergic rhinitis and immunotherapy patients: a five-year retrospective analysis and web-based implementation [abstract]. *Ann Allergy Asthma Immunol* 2008; 100.
- 23 Dalan D. E-mail & Internet-based allergic rhinitis survey showing improved adherence & outcomes [abstract]. *Ann Allergy Asthma Immunol* 2007; 98.

- 24 Chen E, Stetson P, Lussier Y, *et al.* Detection of practice pattern trends through natural language processing of clinical narratives and biomedical literature. *AMIA Symposium Proceedings Archive* 2007. pp. 120–124.
- 25 Seyfried L, Hanauer D, Nease D, *et al.* Enhanced identification of eligibility for depression research using an electronic medical record search engine. *Int J Med Inform* 2009; 78:e13–e18.
This is a hybrid supervised/human NLP data mining technique that was able to find patients who met clinical trial criteria through a semi-automated analysis of discharge summaries.
- 26 Conway M, Doan S, Kawazoe A. Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inform* 2009; 78:e47–e58.
This is unsupervised exploratory data mining using additional refinements of NLP of online news reports, and they were able to monitor, and then find a new disease outbreak report using this technique.
- 27 Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res* 2009; 11:e11.
The author pioneered real-time data mining and monitoring of influenza and continuing other public health issues online using *Google*.
- 28 Dalan D. Google ad word hits correlated with ragweed pollen counts of the Fargo NAB station. *J Allergy Clin Immunol* (in press).
- 29 Schein C, Ivanciuc O, Braun W. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol Allergy Clin North Am* 2007; 27:1–27.
- 30 Samarghitean C, Vihinen M. Bioinformatics services related to diagnosis of primary immunodeficiency's. *Curr Opin Allergy Clin Immunol* 2009; 9:531–536. This article promotes and recommends the use of many databases that are available now online, facilitated by informatics tools in the evaluation, analysis, and prediction of this disease.
- 31 Gendel S. Allergen databases and allergen semantics. *Regul Toxicol Pharmacol* 2009; 54 (Suppl 1):S7–S10.
The author discusses the differences in content, organization, and accessibility of online databases important in food proteins and allergenicity and promotes a semantic web procedural tool to overcome barriers to promote better use of these registries.
- 32 Sittig DF. Potential impact of advanced clinical information technology on healthcare in 2015. *Stud Health Technol Inform* 2004; 107:1379–1382.
- 33 Dalan D. Boost your asthma IQ. Web-based tool enhances specialists' knowledge of NHLBI guidelines. *ADVANCE for Managers of Respiratory Care*. <http://respiratory-care-sleep-medicine.advanceweb.com/Editorial/Content/Editorial.aspx?CC=123652>.
The author evaluated the use of this online central registry that creates meaningful outcomes reports that can be used for performance measures in asthma.
- 34 Butler J, Norris J, Dalan D, *et al.* Online healthcare: a classic clash of technology, people and processes. In: Einbinder L, Lorenzi NM, Ash JS, Gadd CS, Einbinder J, editors. *Transforming healthcare through information: case studies series: health informatics*. 3rd ed. Seacaucus, New Jersey: Springer-Verlag; 2010.
The authors describe necessary changes for a large clinic implementing EHRs and online healthcare services that required an organizational makeover.
- 35 Lorenz N, Riley R, editors. *Managing technological change: organizational aspects of health informatics*. 2nd ed. Seacaucus, New Jersey: Springer-Verlag; 2004.