system [14] takes the "firehose" approach - automatically ingesting around 1,500 of the most common data elements from inpatient notes and performing association statistics in order to predict next order recommendations and outcomes. Importantly, they found that using temporal relationships between orders in their database improves results, from a precision at 10 recommendations of 33% to 38%.

A team at Mount Sinai has also developed an unsupervised method for learning directly from EHR data, this time using state-of-the-art artificial intelligence (AI) techniques such as feature learning and deep neural networks, called Deep Patient [15]. This system was used to predict whether patients would develop various diseases using random forest classifiers after using a deep neural network for feature extraction. The system was found to outperform other unsupervised learning mechanisms such as Principal Component Analysis (PCA) and Gaussian mixture models. Accuracy of the system was found to be quite high (.929) but the F-Score was still rather low (.181), even though it was better than all comparison systems. These "firehose"-based approaches are sure to continue gaining popularity as more structured and free text EHR data is annotated with standardized semantic resources for input into such systems. This strategy is related to those used in many papers reviewed by the section editors based on the extraction of large number of quantitative features in medical images (i.e. radiomics), and on the use of raw EHR data to build predictors, as in the work of Singh, *et al.* [16], identifying novel predictors of kidney failure from concepts extracted directly from clinical notes.

Acknowledgements
We would like to acknowledge the support of Brigitte Séroussi and John H. Holmes, along with the reviewers who assisted with the selection process.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
2. Ryan PB, Stang PE, Overhage JM, Suchard MA, Hartzema AG, DuMouchel W, et al. A comparison of the empirical performance of methods for a risk identification system. Drug Saf 2013;36 Suppl 1:S143–58.
3. Li G, Sajobi TT, Menon BK, Korngut L, Lowerison M, James M, et al. Registry-based randomized controlled trials- what are the advantages, challenges, and areas for future research? J Clin Epidemiol 2016;80:16–24.
4. Shivade C, Raghavan P, Fosler-Lussier E, , Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221–30.
5. Lamy J-B, Séroussi B, Griffon N, Kerdelhué G, Jaulent MC, Bouaud J. Toward a formalization of the process to select IMIA Yearbook best papers. Methods Inf Med 2015;54:135–44.
6. Sahoo SS, Zhang G-Q, Bamps Y, Fraser R, Stoll S, Lhatoo SD, et al. Managing information well: Toward an ontology-driven informatics platform for data sharing and secondary use in epilepsy self-management research centers. Health Informatics J 2016;22:548–61.
7. Kamdar MR, Tudorache T, Musen MA. A systematic analysis of term reuse and term overlap across biomedical ontologies. Semantic Web 2016;:1–19.
8. Sauer BC, Jones BE, Globe G, Leng J, Lu CC, He T, et al. Performance of a Natural Language Processing (NLP) Tool to Extract Pulmonary Function Test (PFT) Reports from Structured and Semistructured Veteran Affairs (VA) Data. EGEMS (Wash DC) 2016;4:1217.
9. Sáez C, Zurriaga O, Pérez-Panadés J, Melchor I, Robles M, García-Gómez JM, et al. Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories. J Am Med Inform Assoc 2016;23:1085–95.
10. Demner-Fushman D, Kohli MD, Rosenman MB, Melchor I, Robles M, García-Gómez JM. Preparing a collection of radiology examinations for distribution and retrieval. J Am Med Inform Assoc 2016;23:304–10.
11. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
12. Prasser F, Kohlmayer F, Kuhn KA. The Importance of Context: Risk-based De-identification of Biomedical Data. Methods Inf Med 2016;55:347–55.
13. Goldstein BA, Pencina MJ, Montez-Rath ME, Winkelmayer WC. Predicting mortality over different time horizons: which data elements are needed? J Am Med Inform Assoc 2017;24:176–81.
14. Chen JH, Podchiyska T, Altman RB. OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records. J Am Med Inform Assoc 2016;23:339–48.
15. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Sci Rep 2016;6:26094.
16. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A Concept-Wide Association Study of Clinical Notes to Discover New Predictors of Kidney Failure. Clin J Am Soc Nephrol 2016;11:2150–8.

Correspondence to:
Daniel R. Schlegel
Department of Computer Science
396 Shineman Center
SUNY Oswego
Oswego NY, 13126, USA
E-mail: daniel.schlegel@oswego.edu

Summary of the Best Papers Selected for the 2017 Edition of the IMIA Yearbook, Special Section "Learning from Experience: Secondary Use of Patient Data"

Chen J, Podchiyska T, Altman R

OrderRex: clinical order decision support and outcome predictions by data-mining electronic medical records

J Am Med Inform Assoc 2016;23:339-48

Compliance with evidence-based guidelines is low and a majority of clinical decisions are not supported by randomized control trials. Thus, a large part of medical practice is thus driven by individual expert opinion. The authors present a clinical order recommender system which operates on a database which has been mined from existing patient data. The input to the data mining system is around 1,500 common electronic medical record (EMR) data elements (out of 5.4 million structured data elements) from labs results, orders, and diagnosis codes, including temporal separation in the form of patient timelines. This data was extracted for 18 thousand patients and stored in an association matrix. Queries to the database come in the form of clinical terms for the captured data elements for a patient. A ranking of suggested orders based on the input data and the association matrix is output to the user. By mixing outcomes such as death and hospital readmission in with the order results, the system also acts as a predictor of outcomes. The authors observe that including the temporal data increased precision from 33 to 38%, but also note that continued work is required to differentiate simply common behaviors on certain data from the correct ones.

Miotto R, Li L, Kidd BA, Dudley JT
Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records
Sci Rep 2016;6:26094

Proposed in this paper is a novel unsupervised deep feature learning method to derive a patient representation from EHR data that facilitates the prediction of clinical outcomes. Deep learning techniques, using neural networks with more than one hidden layer, have not previously been broadly used with EHR data. The authors used aggregated medical records from the Mount Sinai data warehouse with a stack of denoising auto-encoders to capture stable structures and regular patterns from pre-processed EHR data. Then, they implemented random forest classifiers (one-vs.-all learning) to predict the probability that patients might develop a certain disease. On 76,214 test patients comprising 78 diseases from diverse clinical domains and temporal windows, the results significantly outperformed those achieved using representations based on raw EHR data and alternative feature learning strategies such as principal component analysis and Gaussian mixture models.

Saez C, Zurriaga O, Perez-Panades J, Melchor I, Robles M, Garcia-Gomez JM
Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in Spain: a systematic approach to quality control of repositories
J Am Med Inform Assoc 2016;23:1085-95

The authors propose the evaluation of variability in data distributions as a criterion which could be used systematically in assessing data quality. This variability is assessed first on different sources of data (i.e., from different sites), and second, over time. The authors proposed a novel statistics-based assessment method providing data quality metrics and exploratory visualizations. The method is empirically driven on a public health mortality registry of the region of Valencia, Spain, with >500,000 entries from 2000 to 2012, separated into 24 health departments. The repository was partitioned into two temporal subgroups following a change in the Spanish National Date certificate in 2009. Several types of data quality issues were identified including punctual temporal anomalies, and outlying or clustered health departments. The authors note that these issues can occur because of biases in practice, different populations, and changes in protocols or guidelines over time - none of which are solved through usual techniques of mapping to standard semantics.

Prasser F, Kohlmayer F, Kuhn KA
The Importance of Context: Risk-based De-identification of Biomedical Data
Methods Inf Med 2016;55:347-55

As data sharing becomes more common, concerns about maintaining the privacy of patients in such data sets is growing as well. International laws, such as HIPAA, and European Directive on Data Protection emphasize the importance of context when implementing measures for data protection. With methods of de-identification such as k-anonymity (dataset is transformed in such a way that each record is not different from k-1 other records), the degree of protection is high, but it is associated with a loss of information content. Indeed, a major challenge of data sharing is the adequate balance between data quality and privacy. The authors propose a generic de-identification method based on risk models, which assesses the risk of re-identification. An experimental evaluation was performed to assess the impact of different risk models and assumptions about the background knowledge/context of an attacker. Compared with reference methods, the loss of information was between 10% and 24% less, depending on the strength of the adversary being protected against.