Biopsy Records Do Not Reduce Diagnosis Variability in Cancer Patient EHRs: Are We More Uncertain After Knowing?

Jose-Franck Diaz-Garelli, Ph.D., Brian J. Wells, M.D. Ph.D., Caleb Yelton, Roy Strowd, M.D., Umit Topaloglu, Ph.D.
Wake Forest Baptist Medical Center, Winston Salem, NC

Abstract

Diagnostic codes are crucial for analyses of electronic health record (EHR) data but their accuracy and precision are often lacking. Although providers enter precise diagnoses into progress notes, billing standards may limit the particularity of a diagnostic code. Variability also arises from the creation of multiple descriptions for a particular diagnostic code. We hypothesized that the variability of diagnostic codes would be greater before surgical pathology results were recorded in the medical record. A well annotated cohort of patients with brain neoplasms was studied. After diagnostic pathology reporting, the odds of more distinct diagnostic descriptions were 2.30 times higher (p=0.00358), entropy in diagnostic sequences was 2.26 times higher (p=0.0259) and entropy in diagnostic precision scores was 15.5 times higher (p=0.0324). Although diagnostic codes became more distinct on average after diagnostic pathology reporting, there was a paradoxical increase in the variability of the codes selected. Researchers must be aware of the inconsistencies and variability in particularity in structured diagnostic coding despite the presence of a definitive diagnosis.

Introduction

The secondary use of Electronic Health Record (EHR) data is an essential resource for biomedical research. Currently, a plethora of analysis and computing methods have the capacity to identify new patterns and provide new insights from data produced across the translational science spectrum. As new discoveries are introduced, feeding them back into clinical practice will build learning healthcare systems that promise to improve population health, lower health care costs, empower consumers, and facilitate innovation.[1] Consequently, EHR data will increasingly become a valuable resource for a number of research fields, including Comparative Effectiveness Research, Precision Genomic Medicine, Patient Reported Outcomes, etc.[2] Health services researchers, for example, routinely use clinical data to carry out a wide variety of statistical and epidemiological studies to uncover actionable information to make healthcare delivery systems more efficient and effective.[3] These analyses are often supported by clinical data that resides in a data warehouse/repository for longitudinal, retrospective and other types of studies.[4–7] Research data warehouses have been implemented as the platform that will allow researchers to identify cohorts and extract data before and/or after Institutional Review Board approval is obtained.[8] Diagnoses are usually the entry point to identify a patient set in data warehouses. Therefore, the accurate assignment of diagnosis (DX) codes is crucial to the successful data reuse of clinical data for evidence based medicine and towards a functioning leaning healthcare system.

Furthermore, timely, complete and accurate diagnosis is critical to support the decisions on a treatment regimen modality in many clinical oncology practices (e.g. surgery, radiotherapy, chemotherapy, or some combination of these),[9,10] clinical symptoms are indicative and supportive for a definitive diagnosis, however, a surgical pathologist's tumor diagnosis is essential to establish an adequate level of certainty.[11–13] One of the main DX coding systems to hierarchically represent all known diagnoses and symptoms is the International Statistical Classification of Diseases and Related Health Problems (ICD), which is provided and maintained by the World Health Organization (WHO).[14] In its 10th revision of the nosology, the ICD-10 classification is routinely employed in healthcare practices via its ICD-10 Clinical Modification (CM) form in the United States[15] and the WHO's version globally. ICDs are used extensively in research, reimbursement, policymaking[16] and its 11th version (i.e. ICD-11) anticipated to be released in 2018.

Due to ever growing understanding of diseases and related delineation challenges, diagnosis accuracy has been studied for decades and early studies have identified large numbers of errors in diagnostic and procedure code assignment.[17–21] Despite its usefulness, ICD's complex hierarchical nature tends to induce concordance errors between physician and billing specialist code selection. As newer versions of ICD are released, error rates have dropped (e.g. from 20-70% in the 1970s to 20% in 1980s).[22] Yet, ambiguity in the ICD-10 hierarchy remains; for instance, Malignant neoplasm of brain (C71) and its child in the hierarchy, Malignant neoplasm of brain unspecified

(C71.9) refers to the same concept. Nonetheless, research regarding these errors shows that the ICD classification structure and the quality of DX recording are both to blame.[23,24] Furthermore, it has been suggested that additional clinical data should be used to identify patients for secondary analysis of data beyond DX codes due to their inaccuracies.[25] Still, the focus of past research has been on the coding process, error classifications, identifying frequent coding errors and defining preventive measures and minimize the impact of these errors.

Despite these well-known challenges to the accurate and systematic recording of DX, existing EHR system features further contribute to diagnostic variability. Many EHRs provide multiple DX descriptions for each individual DX code.[26] Such features intend to allow clinicians to select the most appropriate textual description. However, this additional complexity may have important unintended consequences. For example, (1) providing more DX descriptions increases the potential for inaccurate DX selection, as compared to more compact terminologies such as ICD-10; (2) giving users multiple descriptions for a single DX code increases the potential DX description variability within a patient's chart and (3) given the broadness and variability in textual DX descriptions, it is likely that these descriptions have varying levels of semantic precision. The first consequence is the amplification of the issues discussed in the literature,[23,24] whereas the latter two are proper to this new mode of DX code selection and, to our knowledge, have not been studied in the literature.

To explore the impact of this setup on the DX recording in EHR systems, we investigated the variability of DX descriptions recorded for a group of patients with a specific and well-defined disease. We also investigated the variability in the particularity (i.e., how much semantic information is included in the description to differentiate the DX) of these DX descriptions as a measure of variation in the semantic specificity of the logged DX. For this analysis brain neoplasms were selected because a large annotated dataset was available to corroborate secondary data extraction, many textual diagnosis descriptions exist for a limited list of specific diagnosis codes, and definitive histopathology is available for all patients to explore changes in diagnosis variability before and after this gold standard diagnosis description is made available in the patient's chart. Specifically, we investigated the differences in patient encounter's primary DX of brain neoplasms (i.e., ICD-10 diagnosis code, C71.*) before and after a Diagnostic Biopsy (BX) report is received. We explore three hypotheses: (I) There is a difference between the number of distinct DX relating to the brain neoplasm condition before and after the BX is recorded (i.e., the variability range is different), (II) The degree of "particularity" (i.e., how much semantic information is included in the description to differentiate the DX) of a DX is higher after the BX, and (III) The variability of this degree of "particularity" is different before and after the BX is received. Our overarching hypothesis is that the presence of presumably true information in the EHR does not guarantee more concordant or particular recording of structured EHR data to document care workflows. We present descriptive and summary statistics of the population used for this study and then test these hypotheses using statistical modelling. This exploratory analysis contributes to the current understanding of DX logging in EHR systems to support the development of a reliable learning healthcare system and evidence-based medicine.

Methods

Upon obtaining the approval from the Wake Forest University School of Medicine's Institutional Review Board (IRB), the encounter diagnosis codes, diagnosis names, encounter dates, ICD-10 codes for the encounter, surgical pathology reports were extracted from the Wake Forest Baptist Medical Center's Translational Data Warehouse. We've employed a combination of statistical and natural language processing (NLP) tools to test our hypotheses. We built binomial regressions to predict whether DX description data belonged to a pre-BX or post-BX record. Specifically, we investigated the difference in number of distinct DX descriptions as well as the degree of DX description particularity using mixed models to account for intra-subject correlation. This second analysis was based on a "particularity score" derived from clinical concepts extracted from these EHR system-specific textual DX descriptions using NLP tools. Lastly, we have investigated particularity score entropy (i.e. a measure of diversity representing the average amount of information in the sequences of scores), as a measure of variability in the DX descriptions chosen by users before and after the BX.

To understand the impact of BX recoding on the charting of brain neoplasm DX and compare our findings, we've utilized 36 patient primary DX data sets, which were extracted during a previous chart review of patients with a brain neoplasm diagnoses as a "gold standard". Comprehensive medical record review was performed of each patient by two independent reviewers. The primary post-operative diagnosis was determined based on review of clinician notes. All treating clinicians were available for consultation when needed. Discrepancies between the two reviewers were resolved by an independent neuro-oncologist. Our final analytical dataset was extracted from the

Translational Data Warehouse and contained 1,385 primary encounter DX observations of 31 patients, recorded from January 1st, 2015 to August 31st, 2017. This time frame was defined to ensure ICD coding version consistency (i.e., to include DXs after October 2015; ICD-10 implementation date). Four patients from the initial date did not have any neoplasm DX or BX data within the selected time window. One patient was excluded due to a confirmed neurofibromatosis DX, which would make the patient's neoplasm DX timeline much more complex and not clinically comparable to other patients in the set.

Our initial dataset consisted of DX records with their corresponding timestamp and patient identifier. Each DX had a specific DX description and was associated with an ICD-10 code. Each patients' initial BX result recording date was added to each DX record with the 'After BX' indicator that served at the dichotomous main outcome variable for our binomial regressions. Additionally, the summary statistics such as mean, median and extreme values were employed to screen the data for outliers, missing values and erroneous input. Dates were also reviewed for potential errors such as values being outside the study's time window. We verified the normality of continuous variables using histograms.

To augment our dataset with a measure of each DX description's particularity, we first created a DX particularity scoring system based on clinical concepts extracted using an NLP tool. We generated particularity scores for each DX description in our dataset. These DX descriptions are part of the EHR system and aim to provide clinicians with clinically-relevant and specific labels to facilitate the selection of DX codes on the clinical practice side (i.e., each DX description was linked to an ICD-10 code, C71.* codes in our dataset). We extracted all medical concepts mentioned in these descriptions using NOBLE Coder[27], an NLP named entity recognition tool for biomedical text, based on the NCI thesaurus terminology[28]. Extracted concepts were of three kinds: Neoplastic processes, Body locations[28] and other concepts that did not increase particularity (i.e., non-diagnostic finding, procedure, functional and non-specific spatial concepts). We scored each DX according based on these concepts by 1) classifying the neoplastic process concepts according to their "depth" in the NCI thesaurus[28] concept hierarchy using score "0" for the most general neoplastic process concept found and adding one score point per child concept separating them highest level (e.g., glioma was one of the most general neoplastic processes and had a score of "0", whereas more specific concepts such as glioblastoma scored "1" as a direct child of the glioma concept) and 2) adding a score point for a body part/organ or tissue concept present in the description or when the DX description corresponded to a location-specific ICD-10 code such as C71.2 (i.e., Malignant neoplasm of temporal lobe). The scores for all concepts extracted per DX were added up, which provided a particularity score for each DX label appearing in the dataset. As a result, we had DX descriptions had particularity degrees ranging from "Malignant neoplasm of brain, unspecified location" to "Oligoastrocytoma of frontal lobe" from least to most particular (scored 0 and 3, respectively). Finally, we calculated the entropy (i.e. a measure of diversity representing the average amount of information in the sequences of scores) of each DX particularity score sequence and DX description sequence for each patient record before and after DX, using the 'entropy' R package.[29] This measure served to provide additional insights into the variability of DX particularity around the around the BX date. *More specifically, entropy results served as a measure of DX particularity score variation and DX description diversity.*

To test our hypotheses, we built binomial regressions using R's generalized linear model (GLM)[30] and mixed models (lme4) packages.[31] We selected binomial regressions because our main outcome variable carried across regressions, 'After BX' was dichotomous. It described whether each DX description group was recorded before or after the BX results were recorded in the EHR. This variable was selected because all our hypotheses aimed to explore differences in counts, continuous values and other derived features before and after the BX. To evaluate hypothesis (I) on the differences in distinct DX counts, we built a model predicting the 'After DX' variable based on the number of distinct DX descriptions. We used a binomial mixed model regression to assess the relationship between DX particularity scores before and after BX (i.e. hypothesis II). We chose a mixed model to account for each DX score as an independent test and subjects as independent from each other, but also be able account for intra-subject correlation; we attributed random intercepts to each patient. We used a GLM binomial regression model to explore relationships between the 'After BX' dichotomous variable and variables derived from the particularity score (i.e., mean, median and standard deviation across each patient's record). Hypothesis III was tested with an additional GLM binomial regression to evaluate differences in entropy[32] of DX particularity scores before and after BX within each patient record. We re-ran each of these regressions using a time window of 90 days before and after the BX to confirm the effect's robustness, eliminate potential censorship and temporal biases. We tested for model improvement by the including covariates such as DX recording time, the max number of days before and

after the BX per patient, the number of distinct clinicians recording DX codes, the number of total DX per patient and the number of departments associated with the encounters. We also tested for variable interactions in all models with more than one variable.

Multiple software tools were used to carry out this analysis. Data massaging was done using a DataGrip software client (version 2017.2.2, JetBrains s.r.o., Prague, Czech Republic). Visual exploration and analyses were done using Tableau (version 10.2.4, Tableau Software, Inc., Seattle, WA). All statistical analyses and data manipulation such as data scrubbing and reshaping were done in R version 3.4.1[30] and RStudio (version 1.0.136, RStudio, Inc., Boston, MA); Statistical significance was set at p=0.05 for all models.

Results

The final analytical dataset included contained 1,385 DX recordings for 31 patients out of which 19 had a total of 186 C71.* DX recordings before their first biopsy dates (Table 1). Only the site-specific ICD-10 C71.0 through C71.4 appeared in the dataset but most DX were associated with C-71.9, Malignant neoplasm of brain, unspecified (73.6% overall, 84.4% before BX and 71.9% after BX); Only C71.1 and C71.2 were present before biopsy besides C71.9. 34 standardized DX descriptions were associated with this 6 ICD-10 codes overall with only 21 appearing before biopsy. 8 distinct providers appeared in the dataset from 31 distinct encounter departments. The number of days before and after biopsy oscillated between 881 days after BX and 780 days before BX, averaging 180±264 days. The particularity score attributed to each DX oscillated between 0 and 3 with averages of 1.25±0.60 overall, 1.20±0.47 before BX and 1.26±0.62 after BX.

Table 1 – Descriptive statistics

Measure	*Overall*	*Before BX*	*After BX*
Distinct Patients	31	19	31
Number of DX Records	1385	186	1,199
Distinct ICD-10 Codes	6	3	6
Distinct DX Descriptions	34	21	32
Distinct Providers	8	6	8
Distinct Hospital Department	31	13	29
Days from Biopsy (Mean±Std.Dev.)	180±264	254±222	247±197
DX Particularity Score (Mean±Std.Dev.)	1.25±0.60	1.20±0.47	1.26±0.62

The difference between the number of distinct DX before and after BX was clearly shown by our first binomial regression model returning an odds ratio of 2.30 (β=0.833, p=0.003). This confirms hypothesis I, showing that with an increase of one distinct DX, a patient's DX list would be 2.3 times more likely to be recorded after the BX (Table 2). This is not to be confused with the number of visits and correspondingly number of primary diagnosis pre- and post-BX. Covariates such as number of distinct clinicians and distinct number of departments were included in preliminary models but did not show a significant effect on whether the distinct DX list occurred after the BX. The maximum number of days from BX was significant but was not included in either model because of its very small coefficient (β=0.00082, OR=1.00082, p= 0.0165). The same regression models for data windowed 90 days before and after the BX date yielded the same results.

Table 2 – Number of Distinct DX Regression

Model Type	*Term*	*Estimate (β)*	*Ratio (exp(β))*	*Std. Error*	*Ratio Confidence Interval (95%)*		*p-value*
Binomial GLM	Intercept	-1.59	0.203	0.722	0.0437	0.770	0.0271
	Number of Distinct DX	0.833	2.30	0.286	1.40	4.26	0.00358

We were unable to find a statistically significant relationship between pre-post biopsy timing and the DX particularity scores or any derivate measure and were not able to confirm hypothesis II. Particularity scores varied between 0 and 3 with a mean of 1.25±0.60. We derived two additional variables from this score: the difference to each patient's maximum score and the difference to the maximum score after BX. Our mixed model regressions revealed no significant relationships between the pre-post BX indicator or the time variable and this score or its derivate variables. The resulting binomial mixed models yielded predicted effects in the expected direction (i.e. DX scores are more specific after BX on average) but no statistical significance was reached for any of the models. We also found that a DX with a one point higher particularity score would be 64% more likely to be recorded after BX (β=0.495, OR=1.64, p= 0.062) at the lowest p-value (Table 3). None of the explored covariates (e.g. number of DX, number of distinct provides and number of distinct departments) showed any statistically significant relationship, nor improved the fit for the DX particularity score. The regression windowed 90 around the BX yielded the same results.

Table 3 – DX Particularity Score Regression

Model Type	*Term*	*Estimate (ß)*	*Ratio (exp(ß))*	*Std. Error*	*Ratio Confidence Interval (95%)*		*p-value*
Binomial Mixed Model	Intercept	3	20.1	0.728	5.26	104.6	0.0000363
	Particularity Score	0.495	1.64	0.265	0.968	2.74	0.062

We found differences in the standard deviation and entropy of the DX particularity score as well as the entropy of DX sequences before and after the BX. This confirms hypothesis III. Predicting our 'After BX' binary from particularity score's standard deviation with a binary GLM model, showed that for an increase of 1 unit, the DX description sequence would be 28% more likely to be part of the post-BX timeframe (p=0.00394) (Table 4). Our binomial GLM regression on particularity score entropy measures showed that for an increase of 1 unit in the DX particularity's score entropy value, the odds of such particularity score sequence belonging to a post-BX recording would be 15.5 times those of being a recorded in pre-BX (p=0.00394), controlling for the number of distinct providers logging the sequence. Similarly, our binomial GLM regression on DX description sequence entropy values showed that for an increase of 1 unit in the DX sequence entropy value, the odds of such sequence belonging to a post-BX recording would be 9.58 of the odds of being a recorded in pre-BX (p=0.0259), controlling for the number of distinct providers logging the sequence. The effect of the number of distinct DX-logging providers was similar for both regressions (OR=1.26, p=0.00127 and OR=1.23, p=0.00769), suggesting that the odds of having a post-DX sequence, for an increase of one DX-logging provider are 26% and 23% higher in each case, respectively. We explored the inclusion of other covariates such as max number of days, number of distinct DX descriptions and number of distinct departments showed improved the model's predictive power, nor improved model fit. The regression windowed 90 around the BX yielded the same results.

Table 4 –Variability Regressions

Model Type	*Term*	*Estimate (ß)*	*Ratio (exp(ß))*	*Std. Error*	*Ratio Confidence Interval (95%)*		*p-value*
Binomial GLM	Intercept	0.137	1.15	0.065	1.01	1.30	3.97E-02
	Std. Dev. (Particularity Score)	0.25	1.28	0.0825	1.09	1.51	0.00394
Binomial GLM	Intercept	-2.32	0.0983	0.626	0.0124	0.29	0.000202
	Entropy (Particularity Score)	2.74	15.5	1.28	1.42	242.3	0.0324
	Distinct Provider Count	0.228	1.26	0.0707	1.11	1.47	0.00127
Binomial GLM	Intercept	-2.42	0.0889	0.65	0.02	0.273	0.000192
	Entropy (DX Sequence)	2.26	9.58	1.02	1.53	90.7	0.0259
	Distinct Provider Count	0.203	1.23	0.0762	1.07	1.45	0.00769

Discussion

We used statistical regressions to evaluate the differences in DX sequences before and after the BX in records representing patients with confirmed brain neoplasm diagnoses. We found that: (1) The number of distinct brain neoplasm DX attributed to the patient was statistically larger after the BX (hypothesis I), (2) Although, our particularity score failed to show a statistically significant relationship to the 'After BX' variable, the regressions hinted at more particular DX descriptions after the BX (hypothesis II) and (3) The variability of DX particularity scores along with their entropy and DX description sequences are higher after the BX (hypothesis III). The results also show a dependency on the number of providers involved in the DX recording process. Our results support the validity of our overarching hypothesis, showing that the presence of biopsy information (i.e. presumably true information) does not guarantee a more particular and concordant recording of DX codes within electronic patient record. In fact, we were able to show that the variability of DX recording was much higher after the BX.

Our study extends the existing literature by exploring aspects beyond the accuracy of DX codes. Most previous work has focused on evaluating the accuracy of DX code charting [2,16,33] rather than understanding the evolution of the recording. We have considered the DX descriptions and logging patterns with the focus on particularity and variability while the accurate diagnosis for a patient and their specific condition is available in the EHR (i.e. BX report). We were unable to find other studies using or describing our method to measure DX description particularity using NLP methods paired to the NCI Thesaurus Classification. However, our findings are congruent with previous studies.[34,35] We also studied the variability of DX codes over time and found congruent results with the existing literature.[33–36] We conceived this as a proxy to concordance, an indirect indicator of data quality.[37–39] Per Kahn et al.'s definition of the plausibility and concordance data quality dimension[38,39] and our findings, EHR systems should enforce DX entries to be align with pathological findings in cancer patients to avoid unnecessary and inaccurate fluctuations over time.

On the practical side, our results show a challenge to the secondary use of the clinical data and the difficulties of reliably harvesting the most accurate and semantically rich information available. The current DX code logging schemes undermine cohort selection by increasing variability, and in turn uncertainty. This can potentially cause incorrect inclusion of patients into analytical cohorts, even when the correct information is available elsewhere in the chart. The fluctuation and inconsistency of DX codes can be attributed to a wide array of factors including billing code reporting requirements. It is possible that clinicians may be choosing DX codes to facilitate downstream processing rather than documenting care with the highest degree of precision.[40] Another potential pitfall in EHR systems is presenting clinicians with inconsistent DX code selection options for distinct clinical workflows. This effect has been noted in existing literature and can go to the extent of threatening patient safety[41,42]. Lastly, a potentially larger issue is the ability to change DX descriptions using free-text labels for any code. This increases the possibility for further variations and uncontrolled vocabulary instance inclusions. One could argue that the most particular DX is in the clinical progress note, yet to this date it remains challenging to reliably access this information programmatically or for a large number of patients.[43,44] During such attempts, any precision could be lost to knowledge extraction challenges. This leaves researchers with the only option of using manual extraction employing qualified professionals, which is expensive, slow for large cohorts and even unfeasible in some cases. The literature often cites phenotyping,[34,36,45] the use of complex algorithms[34,46] and other technological solutions[47–49] to address these problems but these technologies are still in development. Even though, the challenges around the ICD-10 are well studied including the National Academies of Medicine (formerly Institute of Medicine) and there are frequent calls for a new taxonomy for nosology,[50] the variability issue lies deeper on the EHR system side. We believe that future EHR improvements to support logging consistency and a reimbursement coding workflow that allows clinicians to document to the highest possible particularity level would potentially address these problems.

Our analysis presents three limitations that are mostly related to its preliminary nature. First, we focused our study on a limited population of cancer patients (31 subjects). Some of these subjects met the inclusion criteria but returned no data within the study time window selected to ensure DX coding consistency, which further reduced the set. However, we had a final dataset with over a thousand DX that returned adequate statistical results to test our hypotheses. This focused dataset also ensured a homogeneous patient population that responded to the criterion of having a very specific clinical condition that has its diagnosis dictated by a BX report returned in the EHR. A related limitation is that we did not consider comorbidities, yet the preliminary nature of this study required focus rather than comprehensiveness. Second, our particularity score was defined for this analysis and was not previously validated with a gold standard. Regardless, our DX particularity scoring was simple, transparent and systematic enough that it can be considered a simple analytical task for feature extraction.[51,52] The score was based on summing precise features of anatomical location of the neoplasm and then the degree of precision of the neoplastic

process as a level of depth in the NCI Thesaurus classification,[28] a well-known and standardized classification of clinical concepts. Third, we only evaluated DX description variability for one type of cancer. Given the preliminary nature of this analysis and the lack of other literature in the field covering this topic, we compiled this series of simple statistical analyses showing the phenomenon of higher variability after DX, rather than carry out exhaustive analyses to confirm generalizability to all forms of DX. We do expect, however, to find the same kind of phenomenon based on the authors' past clinical and data reuse experiences. This will be confirmed by future analyses.

Future work will be divided into three segments: Confirmatory analyses to verify the robustness and extent of the variability in DX records, Exploration of root causes and The development of informatics solutions to reduce DX variability, while increasing accuracy. First, we will carry out further analysis to confirm that this variability happens for other types of DX within and outside cancer patient records. Then, we will carry out additional secondary analyses of EHR data to explore potential causes for DX variability such as prescribing habits, billing considerations and insurance claim transaction requirements. Finally, we will employ more robust NLP methods to explore and evaluate the semantic distance between BX reports and DX to further understand the problem, but also to use as a basis for the development of informatics solutions to this DX coding problem.

Conclusion

DX records in cancer patient EHRs are more variable after the biopsy report is recorded. Interventions must be developed and adopted to minimize erratic data recording and automatically increase concordance within the patient record. This will avoid uncertainty, misinterpretations and downstream challenges during data integration and secondary analyses using DX codes. In the era of the learning healthcare systems, the concerns around the quality of the data carries the risk of jeopardizing the successful adoption of the evidence-based care, and thus, introduces challenges around the large-scale studies (i.e. phase 3 or phase 4) that are essential for the successful completion of a clinical research process.[53–55]

Acknowledgements

The work is partially supported by the Cancer Center Support Grant from the National Cancer Institute to the Comprehensive Cancer Center of Wake Forest Baptist Medical Center (P30 CA012197) and by the National Institute of General Medical Sciences' Institutional Research and Academic Career Development Award (IRACDA) program (K12-GM102773). The authors acknowledge use of the services and facilities, funded by the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (UL1TR001420).

References

1. Köpcke, F. & Prokosch, H.-U. Employing Computers for the Recruitment into Clinical Trials: A Comprehensive Systematic Review. *J Med Internet Res* **16,** (2014).
2. Safran, C. Reuse Of Clinical Data. *IMIA Yearbook* **9,** 52–54 (2014).
3. Safran, C. Using routinely collected data for clinical research. *Statistics in Medicine* **10,** 559–564 (1991).
4. Dyer, K., Nichols, J. H., Taylor, M., Miller, R. & Saltz, J. Development of a universal connectivity and data management system. *Crit Care Nurs Q* **24,** 25–38; quiz 2 p following 75 (2001).
5. Szirbik, N. B., Pelletier, C. & Chaussalet, T. Six methodological steps to build medical data warehouses for research. *Int J Med Inform* **75,** 683–691 (2006).
6. Sahama, T. R. & Croll, P. R. A Data Warehouse Architecture for Clinical Data Warehousing. in *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers - Volume 68* 227–232 (Australian Computer Society, Inc., 2007).
7. Lyman, J. A., Scully, K. & Harrison, J. H. The development of health care data warehouses to support data mining. *Clin. Lab. Med.* **28,** 55–71, vi (2008).
8. Schubart, J. R. & Einbinder, J. S. Evaluation of a data warehouse in an academic health sciences center. *International Journal of Medical Informatics* **60,** 319–333 (2000).
9. Boyle, P. & Levin, B. World Cancer Report 2008. *World Cancer Report 2008.* (2008).
10. Soerjomataram, I. *et al.* Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions. *The Lancet* **380,** 1840–1850 (2012).
11. Becker, R. L., Specht, C. S., Jones, R., Rueda-Pedraza, M. E. & O'Leary, T. J. Use of remote video microscopy (telepathology) as an adjunct to neurosurgical frozen section consultation. *Human Pathology* **24,** 909–911 (1993).
12. Fisher, E. S. *et al.* The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *Am J Public Health* **82,** 243–248 (1992).

13. Jollis, J. G. *et al.* Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. *Ann. Intern. Med.* **119,** 844–850 (1993).
14. WHO | International Classification of Diseases. *WHO* Available at: http://www.who.int/classifications/icd/en/. (Accessed: 26th September 2017)
15. ICD - ICD-10-CM - International Classification of Diseases, Tenth Revision, Clinical Modification. (2017). Available at: https://www.cdc.gov/nchs/icd/icd10cm.htm. (Accessed: 26th September 2017)
16. O'Malley, K. J. *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Services Research* **40,** 1620–1639 (2005).
17. Medicine, I. of. *Reliability of Medicare Hospital Discharge Records: Report of a Study*. (1977). doi:10.17226/9930
18. Corn, R. F. The Sensitivity of Prospective Hospital Reimbursement to Errors in Patient Data. *Inquiry* **18,** 351–360 (1981).
19. Doremus, H. D. & Michenzi, E. M. Data Quality: An Illustration of Its Potential Impact upon a Diagnosis-Related Group's Case Mix Index and Reimbursement. *Medical Care* **21,** 1001–1011 (1983).
20. Johnson, A. N. & Appel, G. L. DRGs and Hospital Case Records: Implications for Medicare Case Mix Accuracy. *Inquiry* **21,** 128–134 (1984).
21. Hsia, D. C., Krushat, W. M., Fagan, A. B., Tebbutt, J. A. & Kusserow, R. P. Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-Payment System. *http://dx.doi.org/10.1056/NEJM198802113180604* (2010). doi:10.1056/NEJM198802113180604
22. Lloyd, S. S. & Rissing, J. P. Physician and Coding Errors in Patient Records. *JAMA* **254,** 1330–1336 (1985).
23. Bossuyt, P. M. *et al.* Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Fam Pract* **21,** 4–10 (2004).
24. Green, J. & Wintfeld, N. How Accurate are Hospital Discharge Data for Evaluating Effectiveness of Care? *Medical Care* **31,** 719–731 (1993).
25. Safran, C. & Chute, C. G. Exploration and exploitation of clinical databases. *International Journal of Bio-Medical Computing* **39,** 151–156 (1995).
26. Baskaran, L. N. G. M., Greco, P. J. & Kaelber, D. C. Case Report Medical Eponyms. *Appl Clin Inform* **3,** 349–355 (2012).
27. Tseytlin, E. *et al.* NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics* **17,** (2016).
28. Sioutos, N. *et al.* NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics* **40,** 30–43 (2007).
29. Strimmer, J. H. and K. *entropy: Estimation of Entropy, Mutual Information and Related Quantities*. (2014).
30. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2013).
31. Bates, D. *et al. lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. (2017).
32. Shenkin, P. S., Erman, B. & Mastrandrea, L. D. Information-theoretical entropy as a measure of sequence variability. *Proteins* **11,** 297–313 (1991).
33. Escudié, J.-B. *et al.* A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease. *BMC Medical Informatics and Decision Making* **17,** 140 (2017).
34. Wei, W.-Q. *et al.* Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* **23,** e20–e27 (2016).
35. Pippenger, M., Holloway, R. G. & Vickrey, B. G. Neurologists' use of ICD-9CM codes for dementia. *Neurology* **56,** 1206–1209 (2001).
36. Shivade, C. *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* **21,** 221–230 (2014).
37. Weiskopf, N. G. & Weng, C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association* **20,** 144–151 (2013).
38. Kahn, M. *et al.* Transparent Reporting of Data Quality in Distributed Data Networks. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* **3,** (2015).
39. Kahn, M. G. *et al.* A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC)* **4,** (2016).
40. Richesson, R. L., Horvath, M. M. & Rusincovitch, S. A. Clinical Research Informatics and Electronic Health Record Data. *Yearb Med Inform* **9,** 215–223 (2014).
41. Donaldson, M. S., Corrigan, J. M., Kohn, L. T. & others. *To err is human: building a safer health system*. **6,** (National Academies Press, 2000).
42. Zhang, J. & Walji, M. *Better EHR: usability, workflow and cognitive support in electronic health records*. (2014).
43. Burger, G., Abu-Hanna, A., Keizer, N. de & Cornet, R. Natural language processing in pathology: a scoping review. *Journal of Clinical Pathology* **69,** 949–955 (2016).
44. Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C. & Hurdle, J. F. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 128–144 (2008).
45. Hripcsak, G. & Albers, D. J. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* **20,** 117–121 (2013).
46. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology* **31,** 1102 (2013).

47. Murphy, S. N. *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* **17,** 124–130 (2010).
48. Natter, M. D. *et al.* An i2b2-based, generalizable, open source, self-scaling chronic disease registry. *J Am Med Inform Assoc* **20,** 172–179 (2013).
49. Segagni, D. *et al.* The ONCO-I2b2 project: integrating biobank information and clinical data to support translational research in oncology. *Stud Health Technol Inform* **169,** 887–891 (2011).
50. Council, N. R., Studies, D. on E. and L., Sciences, B. on L. & Disease, C. on A. F. for D. a N. T. of. *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. (National Academies Press, 2011).
51. Scott, S. & Matwin, S. Feature engineering for text classification. in *Proceedings of ICML-99, 16th International Conference on Machine Learning* 379–388 (Morgan Kaufmann Publishers, 1999).
52. Blum, A. L. & Langley, P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* **97,** 245–271 (1997).
53. Lorence, D. Regional variation in medical classification agreement: benchmarking the coding gap. *J Med Syst* **27,** 435–443 (2003).
54. Farzandipour, M., Sheikhtaheri, A. & Sadoughi, F. Effective factors on accuracy of principal diagnosis coding based on International Classification of Diseases, the 10th revision (ICD-10). *International Journal of Information Management* **30,** 78–84 (2010).
55. Burgun, A., Botti, G. & Beux, P. L. Issues in the Design of Medical Ontologies Used for Knowledge Sharing. *Journal of Medical Systems* **25,** 95–108 (2001).