ORIGINAL RESEARCH

Comparative study of disease categories: EHR interrogation

Priscilla O. Okunji*[1], Nawar Shara[2], John Kwagyn[3], Ian Brooks[2], Gina Brown[1], Thomas Mellman[3]

[1] College of Nursing and Allied Health Sciences, Howard University, United States
[2] MedStar Health Research Institute, United States
[3] College of Medicine, Howard University, United States

Received: January 1, 2019 **Accepted:** January 29, 2019 **Online Published:** June 10, 2019
DOI: 10.5430/jnep.v9n9p50 **URL:** https://doi.org/10.5430/jnep.v9n9p50

ABSTRACT

Objective: Although the advancement of electronic health records (EHRs) utilization in clinical research may allow for feasibility studies, and identify patients who are eligible for enrollment in clinical trials, it is a complex process to conduct clinical and translational research studies by merging data from different EMRs. Barriers and challenges such as data interoperability, lack of Health Insurance Portability and Accountability Act (HIPAA) compliant platforms for data integration, and the lack of real efforts to resolve these issues make it harder to conduct these studies. However, it is imperative to note that leveraging EHRs to counterbalance these challenges is an area of intense interest and data sharing from hospitals may enable clinical research with large samples for a moderate or large effect size. To inform this issue, we worked across urban hospitals with data extracted from different systems, for patients diagnosed with diabetes and myocardial infarction, in the year 2013.

Methods: Using ICD 9 codes for diabetes (25,000) and myocardial infarction (41,000), data were extracted from urban hospitals. The data were then cleaned, merged using common fields, and analyzed. It is important to note that ICD 9 was used instead of ICD 10 because one of the hospitals had an already existing dataset extracted from EHR using ICD 9. In addition, the hospital with the ICD 9 dataset provided the original data with the needed variables prior to the grant application that made this project possible.

Results: The result showed that patients discharged in 2013 from the selected urban hospitals with MI, were 3.8 times more likely to die while in admission and 4.2 times in MI+DM patients. However, race and gender were not significant in the adjusted model. Variables that impacted this critical result were age of the patients, followed by low density lipoprotein, systolic blood pressure and body mass index.

Conclusions: Comparative studies for preliminary studies through EHR interrogation is the future. This project has confirmed that similar studies should be encouraged and may lead to preventive health education that may ultimately prevent higher mortality rate in certain population. This project is a proof of concept of how data from different EHR platforms can be used to conduct a comparative study by a direct hospitals EHR interrogation, without additional time needed in bedside data collection or purchase of already collected datasets.

Key Words: Data, Hospitals, Interoperability, Diabetes, Myocardial infarction

1. BACKGROUND

EHRs were initially adopted as the primary data source for observational studies or epidemiological studies or comparative effectiveness studies. The advancement of EHRs utilization to randomized clinical trials may have open the door for study feasibility, to facilitate patient recruitment, and

*Correspondence: Priscilla O. Okunji; Email: priscilla.okunji@Howard.edu; Address: College of Nursing and Allied Health Sciences, Howard University, United States.

the streamlining of data collection at baseline and follow-up. It is worth to note that EHRs data mining has its many challenges with data security and privacy, interoperability of disparate systems and infrastructure maintenance for repeat use of high quality data in clinical research. EHRs house structured data (clinical codes for problems; diagnoses; treatment; and management) as well as unstructured data (free text; images). EHR re-use data are mainly research focused on the structured part of the EHR, thus clinical codes are used to extract meaningful information from the raw EHR data. Healthcare has now developed and met standards and interoperability of its own. While standards grounded in health information exchange may not be the only solution to National and Worldwide health care reform, it certainly does assist in accomplishing the goal. Hence, collaboration between academia, industry, regulatory bodies, policy makers, patients, electronic health record vendors and hospitals is critical for a complete systems and data integration for a meaningful use of electronic health records in answering research questions such as, "Is there any relationship between patient characteristics and outcomes?" Hence, this project described the EHR data extraction and comparative outcomes based on the patients' disease categories.

Health services and policy research require access to complete, accurate, and timely patient and organizational data. Often times, the health-related datasets are created and held by diverse and disparate public, private organizations and individual researchers due to coding challenges. It has been ascertained that the size and complexity of the final set of the codes is determined by a combination of the particular code terminology and the concept of interest. For example, an evaluation of 1,054 code sets created for clinical quality measures found varied sizes from sets containing a single code, to a code set for trauma which included 20,560 codes.[1] The construction and validation of such lists is a non-trivial matter. The creation of sets of clinical codes for querying EHR datasets is a critical important phase of using pre-use data for research. It is usually the initial and a difficult step in research as errors are introduced at this stage of missing or wrongly specified codes could result in selection biases that progresses throughout subsequent analyses which may have a major impact in the research outcome.[2] For instance, rheumatoid arthritis code set differences could induce nearly a sevenfold difference in estimates of the incidence.[2,3] National and international standards are everywhere in our society. Biomedical informatics is the use of data mobilized by nationwide adoption of EHR under the Affordable Care Act (ACA) which aligns well with the National Biomedical Research Act—S.2624-114 by Congress on big data to knowledge (BD2K) as depicted in: `http://www.nhlbi.nih.gov/about/documents/strategic-vision/strategic-goals`.[4] Special efforts are being made by some informatics research project to integrate datasets from disparate systems with different racial and ethnic groups, and researchers from economically, socially, culturally or educationally disadvantaged backgrounds, into informatics careers and perform research that disproportionally affect the underserved.[5]

To overcome this barrier, the investigators believe that extracting meaningful information from separate multiple secondary datasets is needed. Though the authors had some challenges in matching and linking data from disparate systems, this paper provides effectiveness clinical research outcomes between the urban hospitals that would enable best practices sharing to decrease undesirable patients' outcomes in future investigations. In addition, less is known about diabetic myocardial infarction (DMI) than myocardial infarction (MI) generally. Data mining, advanced statistical modeling and predictive data analytics are essential in today's emerging BD2K research which will be used to provide answers to the research questions. Urban hospitals serve the medical needs of patients with diabetes, cardiovascular and other chronic diseases through a multidisciplinary approach to patient care. This study will also identify predictors of the disease categories outcomes by extracting data and harmonizing them to a common data model. The datasets will help identify opportunities for targeted interventions for inpatients at risk for readmission, particularly in groups with disparate health outcomes. Furthermore, the results will provide a foundation for future studies exploring the best practices for systems and data interoperability to answer relevant research questions. To the best of the authors' knowledge, little or no studies on the challenges involving EHR data merge have been investigated among urban hospitals. This project will inform future research activities that encompass data extraction and comparisons of disease categories among inpatients discharged from urban hospitals in 2013.

2. METHOD

Using ICD 9 codes for diabetes (25000) and myocardial infarction (41000), data were extracted from urban hospitals. The data were then cleaned, merged using common fields, and analyzed. Descriptive statistics were used to summarize the data and ANOVA to compare differences in means of specific clinical outcomes between the three disease categories (diabetes mellitus–DM, myocardial infarction–MI, diabetes mellitus and myocardial infarction–DM + MI). Mortality rates due to DM, MI and DM + MI were evaluated and compared using Chi-Square test. Logistic regression analyses were used to examine mortality across the three groups. All the analyses were conducted using IBM SPSS software (version 25.0, IBM SPSS).

2.1 Data cleaning

Normally, data extracted from electronic medical record are notoriously "messy" and require extensive cleaning and cross-checking before they could be confirmed as useful and useable. The most common issue experienced was missing values. The first goal was to compile a metadata file on each of the three categories (diabetes mellitus, diabetes mellitus and myocardial infarction, myocardial infarction). This metadata included information on missingness, as well as other information that were useful both in the analysis as stated in the objective, as well creating a master data management file for sharing with the research community at the end of the study based on appropriate regulations. We have examined data from the urban hospitals disorder M Page (M page is one of the Health IT interventions available to bedside clinicians in the hospital; a browser record embedded in the Cerner interface; these are managed by the hospital's clinical informatics staff and available to the investigators clinical and non-clinical). The M Page collects 83 variables for display for a maximum of two weeks of records (∼150-200 SIRS and disorder alerts). A calendar-like function allows for historic data pulls if needed for validation. Demographics and vital data were complete in most cases. Laboratory and vital sign data were somewhat variable ranging from 70% to 100% and MAP. The M page is limited in its display by the underlying logic of the commands pulling data from the Cerner database and we were confident that missing data in these cases could be pulled from 'nearest neighbor' records in the EHR. Data on lab orders was also variably missing, but again, other copies of orders exist in the record to fill gaps. Data generated from this application allowed further applications to fund advanced remedial training, as well as health IT and human factors research for bedside clinicians to emphasize the timely entry of observations to the EHR to (1) mitigate missingness in essential health IT systems and (2) enhance the functionality of Health IT systems by providing timely and accurate data, especially during the care of critically ill patients.

For serially missing data there were two options: impute or ignore. In serial observations on a close time scale, for example automated blood pressure or respiration measurements at the bedside, simple imputation techniques were used to infer missing data. We inferred missing values in these cases and visualize the serial values using Tableau software to ensure no excess deviation from biological possibility, as well as to ensure trends remain intact. All imputations and our observations of visualized data were recorded in the metadata file. Where it was not possible to impute missing data from closely linked variables we flagged data as missing. Although missing data are always a statistical problem when working with EHR data, we expected this research to gather a large enough volume of patient data that missingness does not skew our output.

A second common issue we faced was found on fields lacking proper automated front end validation. In these cases where there is no fixed range of values to ensure data were entered into the EHR correctly, transcription errors were common. In some cases, these were relatively easy to find, annotate and replace. An example was inverted systolic and diastolic blood pressures, or mixed heart and respiration rates. In other cases, data was annotated or flagged as potentially wrong, and managed statistically at the point of analysis. This included deleting variables that are clearly outside natural biological limits. Again, as with missing data we expected our research to gather a large enough volume of patient data that missingness does not skew our output.

2.2 Limitations

Although secondary data analysis requires extra rigor in data extraction, cleaning and analysis, the investigators were cognizant of the potential benefits that was derived from the proposed study. To overcome any technical and statistical challenges or research design barriers, the investigators were carefully selected and made up of an interdisciplinary team with expertise in biostatistics, biomedical informatics, and healthcare to work on the planning and conduct of this study. Another potential limitation emerged in the difficulty in accessing the institution(s) multiple health and operation systems, and authenticating the hospitals' accuracy in reporting coded procedures and processes. To address this, a good working relationship was established with the key personnel at each of the institutions, despite the fact that not all hospitals were able to provide staff experienced in data science, which delayed our study. Data interoperability of some systems was ensured through the use of standard medical vocabularies such as ICD-9-CM diagnostic coding, CPT procedure coding and drug libraries (LOINC, NDC etc.). Extracted data were mapped to a common data model using open source extract/transform/load tools. All investigators involved in the project were mandated to undergo research integrity training (CITI-basic biomedical certificates), privacy and confidentiality (HIPAA) training.

2.3 Study variable measures

Patient Measures: Age 20-80 years. Gender (dichotomous): male, female. Race (categorical): white, black, other. Other measures include: Body mass index (BMI), Systolic blood pressure (SBP), Diastolic blood pressure (DBP), Low density lipoprotein (LDL), High density lipoprotein (HDL). Outcomes Measures: Mortality (dichotomous) was defined as patient dying in the hospital.

2.4 Statistical analysis

Data were extracted for disease categories to enable comparative outcomes studies within and in between diseases groups. The rational being that there has been controversial discussion on the severity outcomes of DM only, MI only and both with the assumptions that MI category may have more detrimental outcomes than both DM and DMI. Descriptive statistics (frequencies, percentages, mean with standard deviation) as appropriate are used to summarize the data. The ANOVA was used to compare difference in means of specific clinical between the three disease categories- MD, MI, MD+MI. Mortality rates due to DM, MI and DM + MI were evaluated and compared using Chi-Square test. In addition, we performed logistic regression analyses to examine the independence of disease category in predicting mortality in models that adjust for age, gender, and race. All the analyses were conducted using IBM SPSS software (version 25.0, IBM SPSS). All tests were two-sided at the 5% level of significance.

3. Results

We studied 4,350 patients with 2,307 (86.1%) African American, 2,492 (57.3%) females with overall mean age ± SD of 66.2 ± 13.3 years. Table 1 presents the demographic and characteristics of the study patients. African American has the highest percentage of all disease categories, with 86.1% of the diabetic patients, 59.3% of the MI patients and 69.75% of both DM+MI patients.

Table 1. Patient demographic characteristics by disease category. Data are mean ± SD or n (%)

Variable	Diabetes Only (n = 2,680)	MI Only (n = 518)	DM + DMI (n = 1,051)
Gender			
Male	1,593 (59.4%)	286 (46.2%)	485 (52.50%)
Female	1,087 (40.6%)	333 (53.8%)	567 (51.55%)
Age, years	62.48 +/	66.86 +/	67.79 +/
Race			
African American	2,307 (86.1%)	367 (59.3%)	577 (69.75%)
Caucasians	189 (7.1%)	241 (38.9%)	455 (27.3%)
Others	184 (6.8%)	10 (1.6%)	20 (2.8%)

Comparison of the patients variables and outcome mortality showed that age and LDL have the highest significance (p < .000), followed by SBP (.002) and BMI (.012). However, DBP and HDL were not statistically significant at p = .05 (see Table 2).

Table 2. Patients variables by disease categories (DM, MI, DM+MI) and significance

Variables	Disease Category			
	DM	MI	DM ± MI	*p*-value
Age	59.18 ± 14.79	66.12 ± 13.43	67.61 ± 12.28	< .000
BMI	32.05 ± 24.26	29.69 ± 18.92	30.16 ± 15.41	.012
SBP	132.05 ± 21.58	129.45 ± 21.72	129.75 ± 21.22	.002
DBP	72.42 ± 12.41	73.14 ± 12.77	72.95 ± 29.84	.579
LDL	91.27 ± 38.87	92.83 ± 49.03	85.14 ± 41.47	< .000
HDL	44.57 ± 16.70	45.39 ± 16.91	44.02 ± 16.21	.314
Mortality Rate	2.3%	8.9%	10.1%	< .0001

Comparison of mortality rates (MR), showed that overall, there was a significant lower risk of death among DM patients (MR = 2.3%), compared to MI only (MR = 8.9%) and MI + DM patients (MR = 10.1%) (ChiSquare, [df = 2] = 114.3, p < .001) (see Table 2). A logistic regression analyses (see Table 3) that adjust for disease category, race, gender and age, showed that, compared to DM patients, MI patients were 3.8 times more likely to die while in admission (OR

= 3.78, CI = [2.55-5.67]) and 4.2 times in MI+DM patients (OR = 4.21, CI = [2.95-6.01]). Race and gender were not significant in the adjusted model.

Table 3. Logistic regression analyses of the effect of disease category on mortality

Variable	OR	95% CI		Sig.
DM	-	--		-
DM+MI	4.211	2.950	6.012	.000
MI	3.799	2.547	5.668	.000
Gender	1.056	.706	1.402	.706
Race	1.200	.263	1.651	.263
Age	1.033	.000	1.044	.000
Constant	.003	.000		.000

4. Discussion & conclusion

The result of this study has shown that patients discharged in 2013 from these urban hospitals with MI, were 3.8 times more likely to die while in admission and 4.2 times in MI+DM patients while race and gender were not significant in the adjusted model. It is important to note that patient variables that impacted this critical result were age of the patients, followed by low density lipoprotein, systolic blood pressure and body mass index. Although the advancement of EHRs utilization to clinical research may have open the door for study feasibility and the streamlining of data collection at baseline and follow-up. It is worth knowing that there have been concerns regarding some barriers and challenges such as the burdensome, systems interoperability, obstructive data collection, and uncertain generalizability of the results. However, it is imperative to note that leveraging EHRs to counterbalance these trends is an area of intense interest and data sharing from hospitals may optimize more clinical research outcomes with large samples for great effect size. To inform this issue, the investigators have extracted and combined data extracted from urban hospitals for 2013 discharged patients with diabetes and myocardial infarction.

In conclusion, this project is significant because it is an added value to how urban hospital in proximal geographical location could maximize their research potential in population studies through EHR data extraction. In addition, the different extraction methods used in this project could be shared with other investigators interested in doing similar project by interrogating hospitals' EHRs in order to optimize the use of EHRs data for pilot and population studies. Therefore, this study may offer opportunities for local authorities and clinicians to focus on using data extracted from urban hospitals within a regional or geographical proximity, to learn best practices on data mining methodology and usage. In addition, analysis of multisystem datasets from urban hospitals would enable investigators' to specifically answer pertinent research questions that may lead to using best practices among the hospitals. Finally, the results of this project may be used to guide resource allocation to further enhance the data.

Acknowledgements

This project has been funded in whole or in part with Federal funds (UL1TR000101 previously UL1RR031975) from the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through the Clinical and Translational Science Awards Program (CTSA), a trademark of DHHS, part of the Roadmap Initiative, "Re-Engineering the Clinical Research Enterprise".

Conflicts of Interest Disclosure

The authors declare that there is no conflict of interest.

References

[1] Winnenburg R, Bodenreider O. Metrics for assessing the quality of value sets in clinical quality measures, AMIA Annu. Symp. Proc. 2013; 1497-1505.

[2] Nicholson A, Tate AR, Koeling R, et al. What does validation of cases in electronic record databases mean? The potential contribution of free text, Pharmacoepidemiol. Drug Saf. 2011; 321-324. PMid:21351316 https://doi.org/10.1002/pds.2086

[3] Rodríguez LAG, Tolosa LB, Ruigómez A, et al. Rheumatoid arthritis in UK primary care: incidence and prior morbidity, Scand. J. Rheumatol. 2009; 38(3): 173-177. PMid:19117247 https://doi.org/10.1080/03009740802448825

[4] NIH Big Data to Knowledge. 2014. Available from: http://bd2k.nih.gov/#sthash.djnpUibw.14sOg3jy.dpbs

[5] The Kaiser Family Foundation and the American College of Cardiology Foundation. Racial/Ethnic Differences in Cardiac Care: The Weight of the Evidence. (Report #6040) Available from: http://www.kff.org