A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample.

Jonathan P DeShazo

Mark A Hoffman

Follow this and additional works at: https://scholarlyexchange.childrensmercy.org/papers

Part of the Databases and Information Systems Commons, and the Health Services Research Commons

Recommended Citation
DeShazo, J., Hoffman, M. A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample. *BMC health services research [electronic resource]* 15, 384-384 (2015).

This Article is brought to you for free and open access by SHARE @ Children's Mercy. It has been accepted for inclusion in Manuscripts, Articles, Book Chapters and Other Papers by an authorized administrator of SHARE @ Children's Mercy. For more information, please contact library@cmh.edu.
A comparison of a multistate inpatient EHR database to the HCUP Nationwide Inpatient Sample

Jonathan P. DeShazo1* and Mark A. Hoffman2

Abstract

Background: The growing availability of electronic health records (EHRs) in the US could provide researchers with a more detailed and clinically relevant alternative to using claims-based data.

Methods: In this study we compared a very large EHR database (Health Facts©) to a well-established population estimate (Nationwide Inpatient Sample). Weighted comparisons were made using t-value and relative difference over diagnoses and procedures for the year 2010.

Results: The two databases have a similar distribution pattern across all data elements, with 24 of 50 data elements being statistically similar between the two data sources. In general, differences that were found are consistent across diagnosis and procedures categories and were specific to the psychiatric–behavioral and obstetrics–gynecology services areas.

Conclusions: Large EHR databases have the potential to be a useful addition to health services researchers, although they require different analytic techniques compared to administrative databases; more research is needed to understand the differences.

Background

Quality measurement and many health services research studies traditionally rely on administrative claims data. Discrepancies between medical claims-based data and clinical data within the patient medical record are well established [1–3], and claims data has been more consistently available than clinical data. However, recent advances in electronic health record (EHR) and health information exchange technology adoption may increase the availability of clinical data for research.

The Congressional Budget Office estimates that 90 % of physicians and 70 % of hospitals in the US will have electronic health records by 2019 [4]. Although electronic health records are primarily intended to support patient care, there is a high level of interest in using EHR data for secondary purposes such as biomedical and outcomes research, clinical process improvement, and epidemiological monitoring [5–8]. Despite some differences in EHR data quality compared to expert review and self-reported data [9], many health services researchers consider the EHR record to be more accurate compared to claims data and are calling for a shift away from claims-based measures to using clinical measures derived from the EHR [3, 10] for research, quality measurement, and performance monitoring. This transition away from claims data is demonstrated by the recent HITECH Act, which requires providers to report key clinical performance metrics directly from their EHR [11]. Limited information has been published on the benefits and limitations of using EHR data compared to claims-based data [12], yet it is critically important to understand these differences as we begin to rely on EHR data for research and payment in addition to supporting clinical care.

The potential advantages of using EHR data are numerous and are generally related to the detailed nature of the data as well as the benefits of being on a computer-based platform, yet the challenges posed are equally significant. The growing availability of electronic health records in the US could provide researchers with faster, less resource-
intensive access to data, larger population samples, more data measurements, and additional types of data compared with primary data collection methods and claims data [6, 13]. However, researchers have noted significant challenges with using EHR data including privacy issues, variation in EHR data sources, inconsistent case definitions, poor data quality, and questionable representativeness of patient populations [8, 14, 15]. Many of these comparisons are based on public health or research data capture models, in which the data capture instrument and process were designed to capture comprehensive and analysis-ready data, while EHRs are generally implemented to provide clinically effective and rapid workflows with a higher tolerance for missing or absent data.

Notable platform-specific EHR databases that are frequently referenced in publications include the Veterans Health Administration databases [16, 17], the Kaiser Permanente Northern California Research Database [18], and the GE Centricity EMR Database [19]. Although valuable for many types of inquiry, using databases such as these to make population estimates may be problematic due to unique regional variations, inpatient or outpatient bias, or limited demographic population captured in the data.

A large EHR database, however, which is widely dispersed among different populations, hospital types, and care settings, may be used to make estimates of health at the population level or for broader groups such as the national universe of inpatients. Moreover, large EHR databases that contain detailed clinical data not found in claims databases could be very valuable to population health management if it were accurate, consistent, and representative of the population. As large EHR databases grow in numbers and in research utility, it becomes even more important to assess their strengths and limitations compared to existing data sources.

In this study, we compare a very large EHR database (Health Facts®) to a well-established population estimate (Nationwide Inpatient Sample). Although we are primarily concerned with the issue of representativeness, other considerations are relevant to our results. To assess representativeness, we calculate adjusted population estimates for encounter demographics, diagnoses, and procedures from both data sources, as well as the relative differences and t-values between the Nationwide Inpatient Sample and Health Facts® EHR populations.

Cerner health facts®

Cerner Corporation, a leading EHR vendor, maintains one of the largest vendor-specific EHR databases, called Health Facts® (HF). Contributing organizations receive quality and benchmarking reports based on internal and external data. HF data is de-identified and HIPAA-compliant to protect both patient and organization identity. The patient-level data in HF includes encounter, medication, diagnosis, laboratory orders and results, pharmacy, and procedure values. These records are comprehensive and include over 300 data elements. For example encounter information includes details such as payer, discharge and admission sources, and care setting. Laboratory data includes collection sites, reference ranges, and the timing of collection and results. The database is longitudinal and hospital patients can be followed post-discharge if they return to care settings within the same health system. At the time of this writing, HF contains data from more than 500 health care facilities across the United States representing 133 million encounters and 84 million patients over the past two decades. Health Facts® data is a publicly available resource from Cerner Corporation (Kansas City, MO).

AHRRQ HCUP Nationwide inpatient sample

The Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample (NIS) is a valuable resource primarily intended for studies of inpatient hospital utilization and charges (http://www.hcup-us.ahrq.gov/nisoverview.jsp). It is the largest publicly available all-payer inpatient care database for the United States; in 2010, the NIS sampled approximately 20% (1051) of all hospitals in 45 states. National estimates are based on a validated method of weights [20]. The NIS is released annually and contains standardized data elements reporting hospital characteristics, diagnosis and procedure codes, and other important data that can support research on medical treatment and effectiveness, quality of care and patient safety, impact of health policy changes, and use of hospital services within the United States. The research footprint and ultimate value of the HCUP NIS and other HCUP all-payer databases is significant, with thousands of studies publishing on HCUP data in peer-reviewed journals including hundreds of citations in top journals [21].

Methods

Data sources

Both the HF EHR database and the NIS data set capture patient-level data elements linked to inpatient encounters. Both data sources are available to the public at www.cerner.com/lifesciences and https://www.hcup-us.ahrq.gov/nisoverview.jsp respectively.

For this analysis we used an extract from the HF database made in January 2012. To make the HF data comparable to the 2010 NIS data set, we included only acute care discharges made during 2010 and excluded any stay longer than 365 days. We then created a weighting scheme similar to the weighting scheme used in the NIS database to make population estimates. First, we used 2010 American Hospital Association data stratified by region and hospital size to determine how many hospitals and discharges there were in the universe. We then assigned a weight to
each hospital in HF that reflected its representation in the stratification. The formula for weight used is: (# in AHA universe/# in HF for each stratum). Strata used are region and bed size. For example, each 200–299 bed hospital in the Northeast in the HF sample is weighted to be equivalent to 6.23 hospitals in the universe. Because of limitations as a convenience sample, this weighting scheme is not balanced (i.e., we did not try to have the same size groupings) and does not take into account teaching hospitals as the NIS does. We used HCUP NIS data for 2010 for the comparison with no additional modifications.

Analytic methods
Population estimates were calculated using SAS9 (SAS Institute Inc., Cary, NC, USA). Weighted estimates of diagnoses were aggregated using Major Diagnostic Category (MDC) groupings and included primary (i.e., first listed) diagnoses recorded for an encounter in the respective database. Procedures were aggregated according to high-level Clinical Classification Software (CCS) groupings and also included primary (first listed) procedure codes. ICD-9-CM code mappings for MDC and CCS groupings can be found on the HCUP website.

Standard errors were calculated and presented along with the NIS-weighted count estimates. HF provides complete enumeration of discharges from the source; therefore, no standard errors are presented. To compare the two data sets, we use relative difference and present the t-value. Relative difference is represented by the absolute value of (NISest-HFest)/NISest. The t-value is calculated as the difference between the estimations divided by the standard error (NIStot-HFtot)/(NISeststd err). A hypothesis-driven comparison of two data sets consisting of very large data sets such as these would result in statistical significance with very little difference. Therefore, we assess the significance of the t-value at 95 % confidence and, due to such large sample sizes, also at 99.9 % confidence. For those that are significantly different, we also provide the relative difference to assess the magnitude.

Results
Encounters and demographics are presented in Table 1.
At the discharge level, the regional counts of discharges are generally comparable between the NIS data set and the HF EHR population. However, the HF population tends to be slightly younger (by approximately 1 year) and report fewer in-hospital deaths (1.63 vs. 1.90 %) compared to the NIS data set.

Tables 2 and 3 present comparisons on diagnoses and procedures, respectively.

Nearly all of the encounters for both the NIS and HF had primary (i.e., first listed) diagnoses, equaling roughly 39 million encounters for each. In all but two of the MDC disease categories (14, 15), the percentage of the overall proportion of discharges in the NIS data set is within 2 % of the overall proportion of the HF population, indicating that the distribution of disease is very similar between the two data sources. Actual counts of encounters follow a similar pattern on comparability. With 95 % confidence in detecting a difference, 15 of the 25 MDC show statistically significant differences between the data sets. The relative difference is smaller than 5 % in one category, and smaller than 10 % in 5 categories. Although many MDC groups are consistent between the NIS data set and HF population, there are significant differences in the number of primary diagnoses in obstetrics and gynecological diagnoses (MDC 13, 14, 15), as well as psychiatric and behavioral diagnoses (MDC 19, 20).

About two-thirds of the NIS data set and one-half of the HF population had procedure codes associated with the encounter. Despite representing significantly fewer procedures overall, the proportions of procedures done in each CCS

---

Table 1

| Table 1 Encounters and demographics: comparison of nationwide inpatient sample (NIS) and health facts® (HF) EHR population |
|-----------------------------------------------|
| **HCUP NIS** | **Std. error** | **HF** | **Rel diff (%)** | **t-value** |
|-----------------------------------------------|
| Total discharges | 39008298 (100.00 %) | 723539 | 38787005 | 0.57 | 0.31 |
| Northeast | 7579674 (19.43 %) | 317987 | 7603789 (19.60 %) | 0.32 | 0.08 |
| Midwest | 8839256 (22.66 %) | 320253 | 8897944 (22.94 %) | 0.66 | 0.18 |
| South | 14985984 (38.42 %) | 483009 | 15055887 (38.82 %) | 0.47 | 0.14 |
| West | 7603384 (19.49 %) | 294165 | 729385 (18.64 %) | 4.92 | 1.27 |
| Length of Stay (Mean) | 4.7 | 0 | 4.75 | 1.27 |
| Length of Stay (Median) | 3 | N/A | 3 | 1.27 |
| Age (Mean) | 48.36 | 0.37 | 47.38 | 2.03 | 2.65* |
| Male | 16485595 (42.26 %) | 312023 | 16207543 (41.79 %) | 1.69 | 0.89 |
| Female | 22436406 (57.52 %) | 429631 | 22549349 (58.14 %) | 0.50 | 0.26 |
| In-hospital Deaths | 740748 (1.90 %) | 17443 (0.03 %) | 633121 (1.63 %) | 14.53 | 6.17* |

*Indicates t-values over 1.960. These are categories that are significantly different between the data sets at 95 % confidence
category in the HF EHR population were very similar to the proportions in the NIS data set. With 95% confidence, 10 out of 16 NIS CCS categories are different between the data sets. However, only four of the procedure categories (9, 11, 13, 14) had proportions that differed by more than 2% between the NIS data set and the HF population.

**Discussion**

Regarding the demographic data on encounters, the comparison of the NIS data set with the HF EHR population shows strong similarities. The HF EHR population is slightly younger by about a year and reported slightly fewer in-hospital deaths. The small difference in age and mortality could possibly be explained through the distribution of hospitals in the HF database. Compared to the NIS data set, small hospitals (<99 beds) are over-represented in the HF EHR database. If larger hospitals tend to have older patients and increased mortality compared to much smaller hospitals, then this may be reflected in the results. The weighting system applied to HF for this analysis is designed to adjust for much of this variation, yet some difference may remain.

Comparisons of the NIS data set and the HF population for diagnoses and procedures show mixed results. With the exception of a few diagnoses, MDC categories in the HF EHR population were consistent with the NIS data set. Psychiatry and behavioral (Alcohol/Drug Use & Alcohol/Drug Induced Organic Mental Disorders, Mental

**Table 2** Diagnoses: comparison of nationwide inpatient sample (NIS) and health facts® (HF) EHR population

| Major diagnostic category and name | Total number of discharges | Comparison | Percentage of total discharges |
|-----------------------------------|---------------------------|------------|-----------------------------|
|                                   | HCUP | HF | Rel Diff (%) | t-value | HCUP | HF |
| Diseases & disorders of:          |      |    |              |         |      |    |
| Nervous System                    | 2350764 | 58581 | 2373818 | 0.98 | 0.39 | 6.03 | 6.12 |
| The Eye                           | 56970 | 2792 | 52699 | 7.50 | 1.53 | 0.15 | 0.14 |
| The Ear, Nose, Mouth & Throat     | 430252 | 15310 | 453087 | 5.31 | 1.49 | 1.1 | 1.17 |
| The Respiratory System            | 3816282 | 72030 | 3993373 | 4.64 | 2.46* | 9.79 | 10.29 |
| The Circulatory System            | 5315827 | 133686 | 5752792 | 8.22 | 3.27* | 13.64 | 14.82 |
| The Digestive System              | 3473022 | 69682 | 3689595 | 6.22 | 3.10* | 8.91 | 9.51 |
| The Hepatobiliary System & Pancreas | 1147155 | 27534 | 1175248 | 2.45 | 1.02 | 2.94 | 3.03 |
| The Musculoskeletal System & Connective Tissue | 3530758 | 96593 | 3161459 | 10.46 | 3.82* | 9.06 | 8.15 |
| The Skin, Subcutaneous Tissue & Breast | 1007290 | 20998 | 1003193 | 0.41 | 0.20 | 2.58 | 2.58 |
| Endocrine, Nutritional & Metabolic Diseases & Disorders | 1269130 | 28332 | 1338294 | 5.45 | 2.44* | 3.26 | 3.45 |
| The Kidney & Urinary Tract        | 1684021 | 35948 | 1669673 | 0.97 | 0.45 | 4.32 | 4.3 |
| The Male Reproductive System      | 193562 | 8343 | 212178 | 9.62 | 2.23 | 0.5 | 0.55 |
| The Female Reproductive System    | 681145 | 19395 | 214863 | 68.46 | 24.04* | 1.75 | 0.55 |
| Pregnancy, Childbirth & The Puerperium | 4323036 | 145371 | 1609355 | 62.77 | 18.67* | 11.09 | 4.15 |
| Newborns & Other Neonates With Condition Originating In Perinatal Period | 4062084 | 135500 | 3121359 | 23.16 | 6.94* | 10.42 | 8.04 |
| Blood, Blood Forming Organs, & Immunological Disorders | 531030 | 14971 | 526763 | 0.80 | 0.29 | 1.36 | 1.36 |
| Myeloproliferative Diseases & Disorders, Poorly Differentiated Neoplasms | 353805 | 27161 | 334898 | 5.34 | 0.70 | 0.91 | 0.86 |
| Infectious & Parasitic Diseases, Systemic Or Unspecified Sites | 1260596 | 29170 | 1111343 | 11.84 | 5.12* | 3.23 | 2.85 |
| Mental Diseases & Disorders       | 1517367 | 92216 | 862761 | 43.14 | 7.10* | 3.89 | 2.22 |
| Alcohol/Drug Use & Alcohol/Drug Induced Organic Mental Disorders | 484570 | 39402 | 81993 | 62.44 | 7.68* | 1.24 | 0.47 |
| Injuries, Poisonings & Toxic Effects Of Drugs | 609005 | 14033 | 560075 | 8.03 | 3.49* | 1.56 | 1.44 |
| Burns                             | 42428 | 7177 | 46046 | 8.53 | 0.50 | 0.11 | 0.12 |
| Factors Influencing Health Status & Other Contacts With Health Services | 652276 | 31948 | 883098 | 35.39 | 7.22* | 1.67 | 2.28 |
| Multiple Significant Trauma       | 109491 | 8205 | 79620 | 24.11 | 3.08* | 0.27 | 0.21 |
| Human Immunodeficiency Virus Infections | 77981 | 8095 | 49029 | 37.13 | 3.58* | 0.2 | 0.13 |
| All Diagnosis                     | 3897527 | 3880845 |

*Indicates t-values over 1.960. These are diagnoses that are significantly different between the data sets at 95% confidence.
Diseases & Disorders diagnoses as well as obstetrics and gynecology (Diseases & Disorders Of The Female Reproductive System, Pregnancy, Childbirth & The Puerperium, Newborns & Other Neonates With Conditions Originating In Perinatal Period) diagnoses are much less frequently represented in the HF EHR population compared to the NIS data set. It is possible that these differences could be attributed to selection bias of hospitals in the HF database resulting in hospitals that see fewer psychiatry and obstetrics and gynecology cases. For example, hospitals who adopted this specific EHR may see fewer of these kinds of cases. Another, more likely, explanation is that many of those services were still using either paper-based clinical records or a separate EHR system during the study period. This would result in fewer diagnoses in these categories within the hospital EHR. Mental health, substance abuse, and reproductive/sexual health are identified as "sensitive health information" in the recommendations of the National Committee on Vital and Health Statistics to the Secretary of Health and Human Services [22]. Some participating institutions may have compartmentalized this data from the less sensitive EHR data or embraced emerging security features of the EHR that allow patients to have greater control over the privacy of their health data. Information about the frequency of procedures was generally less consistent between the data sources compared to diagnoses. Although 11 of the 15 procedure categories in the NIS data set are proportionally within 2 % of the HF population, the HF population reports many fewer procedures overall compared with the NIS data set. Prior research has identified that EHR databases have tended to capture fewer data elements related to provider orders [23], which may explain the difference. Another possible explanation is that not all contributing sites are able to supply procedure codes to the database.

Differences between claims-based data and EHR data found in our analyses may have implications for quality of care, research design, and policy development. Claims data in general has the advantage of years of validation and research use. Comparably, questions of internal validity and data integrity surround EHR data. Researchers are beginning to address many of these issues by developing innovative data infrastructure and study designs [14, 26]. In contrast, the HCUP NIS is a validated and trusted data source; however, it lacks much of the detailed clinical information that is available through large, longitudinal, detailed EHR databases.

There are overarching considerations to the differences between EHR databases and claims-based samples identified in our findings. First and perhaps primarily, they are different resources created and predominantly used for distinctly different purposes. Prior research has shown claims coding can be markedly different than patient problems captured by providers [1, 24, 25]. Neither claims nor medical records are primarily collected for research so one is not necessarily more 'correct' than the other. Claims data in general has the advantage of years of validation and research use. Comparably, questions of internal validity and data integrity surround EHR data. Researchers are beginning to address many of these issues by developing innovative data infrastructure and study designs [14, 26]. In contrast, the HCUP NIS is a validated and trusted data source; however, it lacks much of the detailed clinical information that is available through large, longitudinal, detailed EHR databases.

### Table 3: Procedures: comparison of nationwide inpatient sample (NIS) and health facts® (HF) EHR population

| CCS Principal Procedure Category and Name | Total number of discharges | Percentage of total discharges | Rel Diff (%) | t-value | HCUP Std. error | HF | HCUP | HF | t-value | HCUP | HF |
|------------------------------------------|----------------------------|--------------------------------|--------------|---------|----------------|----|------|----|---------|------|----|
| Operations on the nervous system         | 7,849,676                  | 24,864                         | 21,309,98    | 8.15    | 1.47           | 2.80 |
| Operations on the endocrine system      | 10,114,181                 | 76,45                          | 57,033       | 43.61   | 5.77*          | 0.19 |
| Operations on the eye                    | 2,643,212                  | 21,212                         | 25,783       | 2.46    | 0.31           | 0.09 |
| Operations on the ear                    | 1,570,613                  | 13,359                         | 25,066       | 59.60   | 7.01*          | 0.08 |
| Operations on the nose; mouth; and pharynx | 1,336,503                 | 88,28                          | 127,031      | 4.95    | 0.76           | 0.43 |
| Operations on the respiratory system    | 65,927,512                 | 26,693                         | 74,993       | 13.75   | 3.40*          | 2.52 |
| Operations on the cardiovascular system | 34,598,612                 | 1,131,346                      | 43,614,95    | 26.06   | 5.96*          | 14.68 |
| Operations on the hemic and lymphatic system | 16,249,312              | 1,179,312                      | 21,861,2     | 34.54   | 4.76*          | 0.74 |
| Operations on the digestive system      | 3,316,780                  | 11,249                         | 31,440,45    | 5.21    | 1.54           | 10.58 |
| Operations on the urinary system        | 5,402,364                  | 35,296                         | 59,594,3     | 10.98   | 1.68           | 2.02 |
| Operations on the male genital organs   | 11,477,052                 | 51,821                         | 70,546       | 38.53   | 8.53*          | 4.66 |
| Operations on the female genital organs | 7,007,975                  | 27,795                         | 79,388,9     | 13.28   | 3.33*          | 2.67 |
| Obstetrical procedures                  | 3,850,461                  | 17,357                         | 39,860,28    | 3.52    | 0.78           | 13.41 |
| Operations on the musculoskeletal system | 29,586,088                | 12,679                         | 22,924,82    | 24.65   | 5.75*          | 7.5 |
| Procedures on the breast                | 7,603,172                  | 32,677                         | 71,733       | 5.65    | 1.32           | 2.41 |
| Miscellaneous diagnostic and therapeutic procedures | 6,025,241                | 35,609                         | 59,701,09    | 5.92    | 6.42*          | 32.3 |
| All Procedures                          | 24,627,198                 | 13,691,84                      | 18,837,298   | 59.28   | 6.42*          | 32.3 |

*Indicates t-values over 1.960. These are procedures that are significantly different between the data sets at 95 % confidence.
Obstetrics or Psychiatry. Significant differences in either each sample. For example, some hospitals may not have
structural or service line differences between the hospitals in the EHR data. It is also possible that there were struc-
tural differences between the data sets in the primary diagno-
sis criteria and definition. There may also be differences
between the groups, and procedure categories in the HCUP NIS data
set and HF EHR population. Compared to the NIS, the
HF EHR population is slightly younger with lower in-
hospital mortality. There tends to be about the same dis-
tribution of principal diagnoses between the groups, with
the HF EHR sample reporting fewer psychiatry, behav-
ioral, obstetrics, and gynecology diagnoses. Compared to
the NIS data set, HF captured fewer procedures overall,
and also in relation to surgeries involving male and female
genital organs and obstetric procedures.

These findings improve our understanding of the dif-
ferences between established claims-based databases and
large EHR databases and support evidence that large
EHR databases have the potential to be a useful addition
to health services researchers.

More research is needed to understand the internal and
external validity of large EHR databases as it pertains
to population health research, specifically in the inter-
pretation of results from EHR databases. There is also
a significant need for the development of innovative
methodologies that may be required to fully utilize these
rapidly growing data sets.

Conclusions
This study compared demographic variables, diagnosis
groups, and procedure categories in the HCUP NIS data
set and HF EHR population. Compared to the NIS, the
HF EHR population is slightly younger with lower in-
hospital mortality. There tends to be about the same dis-
tribution of principal diagnoses between the groups, with
the HF EHR sample reporting fewer psychiatry, behav-
ioral, obstetrics, and gynecology diagnoses. Compared to
the NIS data set, HF captured fewer procedures overall,
and also in relation to surgeries involving male and female
genital organs and obstetric procedures.

These findings improve our understanding of the dif-
ferences between established claims-based databases and
large EHR databases and support evidence that large
EHR databases have the potential to be a useful addition
to health services researchers.

More research is needed to understand the internal and
external validity of large EHR databases as it pertains
to population health research, specifically in the inter-
pretation of results from EHR databases. There is also
a significant need for the development of innovative
methodologies that may be required to fully utilize these
rapidly growing data sets.
Appendix

Table 4 Weighting scheme

| Region   | Bed Size | # of Hospitals in HF Sample | # of Discharges in Sample (Health Facts) | # of Discharges in Universe (AHA) | Discharge Weights (universe/sample) | # of HF Discharges with Dx and Procs | Dx and Proc Weights (universe/sample) |
|----------|----------|----------------------------|-----------------------------------------|----------------------------------|-----------------------------------|-------------------------------------|--------------------------------------|
| Midwest  | <99      | 28                         | 13711                                   | 1226395                          | 89.45                             | 11197                               | 109.53                               |
| Midwest  | 100-199  | 11                         | 49244                                   | 1503133                          | 30.52                             | 18544                               | 81.06                                |
| Midwest  | 200-299  | 5                          | 51911                                   | 1644218                          | 31.67                             | 24916                               | 65.99                                |
| Midwest  | 300+     | 10                         | 152916                                  | 4526492                          | 29.60                             | 59741                               | 75.77                                |
| Northeast| <99      | 23                         | 131065                                  | 385176                           | 2.94                              | 30687                               | 12.55                                |
| Northeast| 100-199  | 7                          | 44618                                   | 1087590                          | 24.38                             | 35543                               | 30.60                                |
| Northeast| 200-299  | 14                         | 230950                                  | 1437848                          | 6.23                              | 74904                               | 19.20                                |
| Northeast| 300-499  | 6                          | 94831                                   | 2021580                          | 21.32                             | 48317                               | 41.84                                |
| Northeast| 500+     | 7                          | 243512                                  | 2689519                          | 11.04                             | 87817                               | 30.63                                |
| South    | <99      | 29                         | 60808                                   | 1632606                          | 27.17                             | 21043                               | 77.58                                |
| South    | 100-199  | 8                          | 69961                                   | 2685597                          | 38.39                             | 26535                               | 101.21                               |
| South    | 200-299  | 9                          | 116086                                  | 2543060                          | 21.91                             | 56292                               | 45.18                                |
| South    | 300-499  | 6                          | 85349                                   | 3766138                          | 44.13                             | 22595                               | 166.68                               |
| South    | 500+     | 6                          | 269477                                  | 4433500                          | 16.45                             | 109494                              | 40.49                                |
| West     | <99      | 10                         | 21522                                   | 778899                           | 36.19                             | 8466                                | 92.00                                |
| West     | 100-199  | 5                          | 32114                                   | 1603107                          | 49.92                             | 2379                                | 673.86                               |
| West     | 200-299  | 4                          | 46321                                   | 1525325                          | 32.93                             | 13427                               | 113.60                               |
| West     | 300+     | 3                          | 80072                                   | 3325308                          | 41.33                             | 22570                               | 147.33                               |

Abbreviations
AHRQ: Agency for healthcare research and quality; CCS: Clinical classifications software (grouping); EHR: Electronic health record; HCUP: Healthcare cost and utilization project; HF: Cerner health facts database; HIPAA: Health insurance portability and accountability act; HITECH: Health information technology for economic and clinical health act; MDC: Major diagnostic category; NIS: National inpatient sample.

Competing interests
JPD declares that he has no competing interests. MH was employed by Cerner during the data acquisition and analysis stages of the research.

Authors’ contributions
JPD conceived of the study, participated in the design of the study, performed the statistical analysis, and drafted the manuscript. MH participated in the design of the study and helped to draft the manuscript. Both authors have read and approved the final manuscript.

Acknowledgements
This research was funded by regular salary support (Virginia Commonwealth University for JPD and Cerner for MH) with no funding from outside institutions. Neither institution was involved in the study design; in the collection, analysis, and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication. The authors thank Anne Renz, who provided editorial services on behalf of Virginia Commonwealth University.

Author details
1Department of Health Administration, School of Allied Health Professions, Virginia Commonwealth University, Grant House Room 201, 1008 East Clay Street, P.O. Box 980203, Richmond, VA 23298-0203, USA. 2Department of Biomedical and Health Informatics, School of Medicine, University of Missouri – Kansas City (UMKC), 2411 Holmes, MG-203B, Kansas City, MO 64108-2792, USA.

Received: 16 February 2014 Accepted: 24 August 2015
Published online: 15 September 2015

References
1. Peabody JW, Luck J, Jain S, Bertenthal D, Glassman P. Assessing the accuracy of administrative data in health information systems. Med Care. 2004;42(11):1066.
2. Benin AL, Vitkauskas G, Thomquist E, Shapiro ED, Concato J, Aslan M, et al. Validity of using an electronic medical record for assessing quality of care in an outpatient setting. Med Care. 2005;43(7):691.
3. Fowles JB, Fowler EJ, Craft C. Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. J Ambul Care Manage. 1998;21(1):24.
4. Steinbrook R. Health care and the American recovery and reinvestment act. N Engl J Med. 2009;360(11):1057–60.
5. Klompas M, McVetta J, Eggleston E, Mendoza T, Daly P, Oppedidano P, et al. Automated surveillance and public health reporting for gestational diabetes incidence and care using electronic health record data. Using administrative databases to identify cases of chronic kidney disease: a systematic review. 2011. p. 23.
6. Kohane IS. Using electronic health records to drive discovery in disease genomics. Nat Rev Genet. 2011;12(6):417–28.
7. Kalra D, Schmidt A, Poets H, Dupont D, Sundgren M, De Moor G. Case report from the EHR4CR project—a european survey on electronic health records systems for clinical research. 2011.
8. Desai JR, Wu P, Nichols GA, Lieu TA, O’Connor PJ. Diabetes and asthma case identification, validation, and representativeness when using electronic health data to construct registries for comparative effectiveness and epidemiologic research. Med Care. 2012;50(S3):1–5. doi:10.1097/MLR.1090b1013e318259c318011.
9. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. Med Care Res Rev. 2010;67(5):503–27.
10. Weiner JP, Fowles JB, Chan KS. New paradigms for measuring clinical performance using electronic health records. Int J Qual Health Care. 2012;24(3):200–5.
11. Health Information Technology for Economic and Clinical Health (HITECH) Act. In Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA). Pub. L. No. 111–5, 123 Stat. 226 (Feb. 17, 2009).
12. Tang PC, Ralston M, Anrigotti MF, Qureshi L, Graham J. Comparison of methodologies for calculating quality measures based on administrative data versus clinical data from an electronic health record system: implications for performance measures. J Am Med Inform Assoc. 2007;14(1):10–5.
13. Blobel B, Hvannberg ET, Gunnarsdótir V. Informatics EFfM, conference EFfMIST, press I: seamless care, safe care: the challenges of interoperability and patient safety in health care: proceedings of the EFMI special topic conference, June 2–4, 2010. Reykjavik, Iceland: IOS Press; 2010.
14. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the american medical informatics association. JAMIA. 2013;20(1):144–51. doi:10.1136/amiajnl-2011-000881.
15. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. J Am Med Inform Assoc. 2012;19(5):684–7. doi:10.1136/amiajnl-2012-000878.
16. Department of Vetrans Affairs. VHA National Patient Care Database (NPCD) [https://catalog.data.gov/dataset/national-patient-care-database-npcd] Accessed 7/20/2015.
17. Hynes DM, Perrin RA, Rappaport S, Stevens JM, Demakis JG. Informatics resources to support health care quality improvement in the veterans health administration. J Am Med Inform Assoc. 2004;11(5):344–50.
18. Terdiman J, Gul J. PS1-39: the kaiser permanente Northern California oracle research database. Clin Med Res. 2011;9(3–4):168–9.
19. Crawford AG, Cote C, Couto J, Daskian M, Gunnansson C, Haas K, et al. Comparison of GE centricity electronic medical record database and national ambulatory medical care survey findings on the prevalence of major conditions in the united states. Population Health Management. 2010;13(3):139–50.
20. Hcup Nationwide Inpatient Sample (NIS): Healthcare Cost and Utilization Project (HCUP). Rockville, MD: Agency for Healthcare Research and Quality; 2007–2009.
21. Schoenman JA, Sutton JP, Kintala S, Love D, Max R. The value of hospital discharge databases. Agency for Healthcare Research and Quality 2005. http://www.hcup-us.ahrq.gov/reports/final_report.pdf
22. Caine K, Hanania R. Patients want granular privacy control over health information in electronic medical records. J Am Med Inform Assoc. 2013;20(1):17–15.
23. Parsonis A, McCullough C, Wang J, Shih S. Validity of electronic health record-derived quality measurement for performance monitoring. J Am Med Inform Assoc. 2012;19(4):604–9. doi:10.1136/amiajnl-2011-000557.
24. Fisher ES, Whaley FS, Krushat WM, Malenka DJ, Fleming C, Baron JA, et al. The accuracy of Medicare’s hospital claims data: progress has been made, but problems remain. Am J Public Health. 1992;82(2):243–8.
25. Jollis JG, Anuszkiewicz M, DeLong ER, Puyor DB, Murphy-Ly, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med. 1993;119(8):844–50.
26. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al. Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data. The SHARPn project. J Biomed Inform. 2012;45(4):763–71. doi:10.1016/j.jbi.2012.01.009.
27. Gold R, Angier H, Mangione-Smith R, Galia C, McIntire PJ, Cowburn S, et al. Feasibility of evaluating the CHIPRA care quality measures in electronic health record data. Pediatrics. 2012;130(1):139–49. doi:10.1542/peds.2011-3705. Epub 2012 Jun 18.
28. Gold R, Angier H, Mangione-Smith R, Galia C, McIntire PJ, Cowburn S, et al. Feasibility of evaluating the CHIPRA care quality measures in electronic health record data. Pediatrics. 2012;130(1):139–49.