Database queries for hospitalizations for acute congestive heart failure: flexible methods and validation based on set theory

Marc Rosenman,[1,2] Jinghua He,[3] Joel Martin,[2] Kavitha Nutakki,[1] George Eckert,[4] Kathleen Lane,[4] Irmina Gradus-Pizlo,[5] Siu L Hui[2,4]

[1]Children's Health Services Research, Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana, USA
[2]Regenstrief Institute, Indianapolis, Indiana, USA
[3]Department of Epidemiology, Merck Sharp & Dohme, North Wales, Pennsylvania, USA
[4]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, Indiana, USA
[5]Krannert Institute of Cardiology, Indiana University School of Medicine, Indianapolis, Indiana, USA

Correspondence to
Dr Marc Rosenman, Children's Health Services Research, Department of Pediatrics, Indiana University School of Medicine, 410 W. 10th Street Suite 1020, Indianapolis, IN 46202, USA; mrosenma@iu.edu

Received 15 April 2013
Revised 20 September 2013
Accepted 23 September 2013
Published Online First
10 October 2013

ABSTRACT

Background and objective Electronic health records databases are increasingly used for identifying cohort populations, covariates, or outcomes, but discerning such clinical 'phenotypes' accurately is an ongoing challenge. We developed a flexible method using overlapping (Venn diagram) queries. Here we describe this approach to find patients hospitalized with acute congestive heart failure (CHF), a sampling strategy for one-by-one 'gold standard' chart review, and calculation of positive predictive value (PPV) and sensitivities, with SEs, across different definitions.

Materials and methods We used retrospective queries of hospitalizations (2002–2011) in the Indiana Network for Patient Care with any CHF ICD-9 diagnoses, a primary diagnosis, an echocardiogram performed, a B-natriuretic peptide (BNP) drawn, or BNP >500 pg/mL. We used a hybrid between proportional sampling by Venn zone and over-sampling non-overlapping zones. The acute CHF (presence/absence) outcome was based on expert chart review using a priori criteria.

Results Among 79 091 hospitalizations, we reviewed 908. A query for *any* ICD-9 code for CHF had PPV 42.8% (SE 1.5%) for acute CHF and sensitivity 94.3% (1.3%). Primary diagnosis of 428 and BNP >500 pg/mL had PPV 90.4% (SE 2.4%) and sensitivity 28.8% (1.1%). PPV was <10% when there was no echocardiogram, no BNP, and no primary diagnosis. 'False positive' hospitalizations were for other heart disease, lung disease, or other reasons.

Conclusions This novel method successfully allowed flexible application and validation of queries for patients hospitalized with acute CHF.

BACKGROUND AND SIGNIFICANCE

Electronic health records (EHRs), including administrative claims and/or medical records databases, are increasingly being used for identifying cohorts with particular clinical characteristics, or 'phenotypes'. These cohort definitions play a role in almost any retrospective analysis of EHR data, as study population, covariate, or outcome. Their greatest value may be realized if they can be developed, validated, and applied efficiently, in order to accelerate the contribution of EHR data to comparative effectiveness and patient-centered outcomes research. Phenotypes are required for genetic association studies,[1 2] drug safety surveillance research,[3–5] and quality of care measurement.[6–8]

However, it takes time to develop phenotype definitions, and then to discern how accurate, useful, and transportable they are.[8–11] The challenges derive from the inherent complexity and uneven quality of real-world EHR data. The data stored in an EHR are shaped by innumerable workflows, interactions, and idiosyncrasies. As Hripcsak and Albers point out, for phenotype development and validation to someday reach a data-driven, high-throughput state, 'the physics of the medical record' will have to be much better understood.[9] How have the data in the EHR been shaped? The difficulties in answering that question—substantial even in one EHR setting—are worse when multi-source data are being used,[12–15] as in a health information exchange (HIE). An HIE receives varying data from many institutions, each of which has its own physics.

The ways that clinical phenotype populations are defined are being scrutinized, in systematic reviews of the literature,[16 17] through comparison of automated results with 'chart reviews' of medical records,[3 18 19] and through application of more involved statistical techniques in drug safety surveillance projects such as OMOP[4 5] and FDA Mini-Sentinel.[3 18] Careful chart review methods are still a mainstay, but the validation literature is more complete for some conditions (myocardial infarction) than for others (osteoporosis). The FDA Mini-Sentinel initiative recently published methods to help address the challenges of conducting such chart reviews across multiple sites, in queries for acute myocardial infarction.[18] The Electronic Medical Records and Genomics (eMERGE) network described lessons learned in validating the transportability of phenotypes across five leading US EHRs.[10] A related challenge is to distinguish an acute exacerbation from the chronic presence of a condition. Queries for vertebral compression fractures had a high positive predictive value (PPV) in identifying patients with a history of such fractures, but it was much more challenging to specifically identify new fractures.[19]

In a leading US HIE—the Indiana Network for Patient Care (INPC)[20]—we developed a flexible method using overlapping (Venn diagram) queries for shaping and validating phenotypes. Here we describe this general approach, in the particular 'use case' to find patients hospitalized with acute congestive heart failure (CHF). We thus faced the challenges of using multi-source HIE data, and of discerning acute versus chronic illness. CHF, as a prevalent and costly chronic disease, is an important clinical domain for EHR queries; electronic data are actively being used in modeling CHF

To cite: Rosenman M, He J, Martin J, *et al. J Am Med Inform Assoc* 2014;**21**: 345–352.

mortality, readmissions, and quality of care.[21–27] Studying CHF also affords the opportunity to use not only traditional billing-type data like International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes but also clinical results like echocardiograms and laboratory values (B-natriuretic peptide (BNP)).[28 29]

We describe the development of the Venn diagram queries for acute CHF hospitalizations, a corresponding sampling strategy for one-by-one 'gold standard' chart review, and statistical methods for calculating PPV and sensitivities, with SEs, across different definitions. We submit that the advantage of this approach is that we do not have to commit to a single set of phenotype criteria at the outset. Rather, one set of chart reviews can be used to validate multiple overlapping criteria for the phenotype definition. Subsequent users of a phenotype query may have various purposes—one size does not fit all.

MATERIALS AND METHODS

Study design and data source

This was a retrospective study of hospitalizations (2002–2011) in the INPC—a leading operational regional HIE. The INPC was formed in 1994 by the Regenstrief Institute and the five major hospital systems in Indianapolis. Its primary purpose is for clinical data exchange, to improve the quality and efficiency of healthcare; it is also used in research. Since 2009, the number of hospital systems has expanded beyond the original five. The INPC employs a master person index, which supports integration of patient data across the different clinical and payer institutions. The availability of data types—for example, diagnoses, procedures, pharmacy transactions, orders, imaging studies, laboratory results, vital signs, and text reports (discharge summaries, radiology reports, etc.)—varies by INPC institution. Because chart review validation would be integral to this study, we selected hospitalizations at two of the largest clinical institutions in the INPC, where ample discharge summary report data would be available for review. The study was approved by the Institutional Review Board of Indiana University.

Inclusion and exclusion criteria

Hospitalizations at the two large clinical institutions were identified, and data from those two institutions or from other institutions including a large healthcare payer—between the dates of hospital admission and discharge—were queried. To be included in any of the phenotype definitions, we required that the patient have had at least one ICD-9 diagnosis for CHF (428.xx, 402.x1, 404.x1, 404.x3, or 398.91) during the hospital stay. To evaluate the sensitivity of the phenotype definitions, we assumed that patients hospitalized with acute CHF who were missed by our inclusion criteria had a high BNP level; therefore, we also queried for hospitalizations in which patients did not have any ICD-9 diagnosis for CHF but did have a maximum BNP >500 pg/mL. Children under age 18 were excluded.

Venn diagram criteria

In consultation with CHF-specialist cardiologists, and based in part on considerations of what data from the INPC would be feasible to analyze in a 1-year project, we selected several circles for the Venn diagram. We made the distinction between the presence of any ICD-9 code for CHF and ICD-9 code 428. We also assessed whether or not the patient had a primary diagnosis of CHF or of ICD-9 code 428. We added queries to delimit the subsets of patients who had an echocardiogram performed during the hospital stay, a BNP drawn during the hospital stay, or a maximum BNP level >500 pg/mL. Complex criteria could then be formed using 'and' and 'or' statements. To help generate a sampling strategy for the 'gold standard' chart review, the number of hospitalizations in each Venn diagram zone (figure 1) was calculated.

Figure 2 shows the Venn diagram zones, each labeled with a letter of the alphabet, to facilitate our reporting of results by different combinations of zones.

In planning for the sampling strategy, we did not consider each Venn zone (or all possible combinations of the zones) to be equally important, but rather focused on 10 sets that, a priori, seemed as if they may have potential utility as phenotypes (table 1).

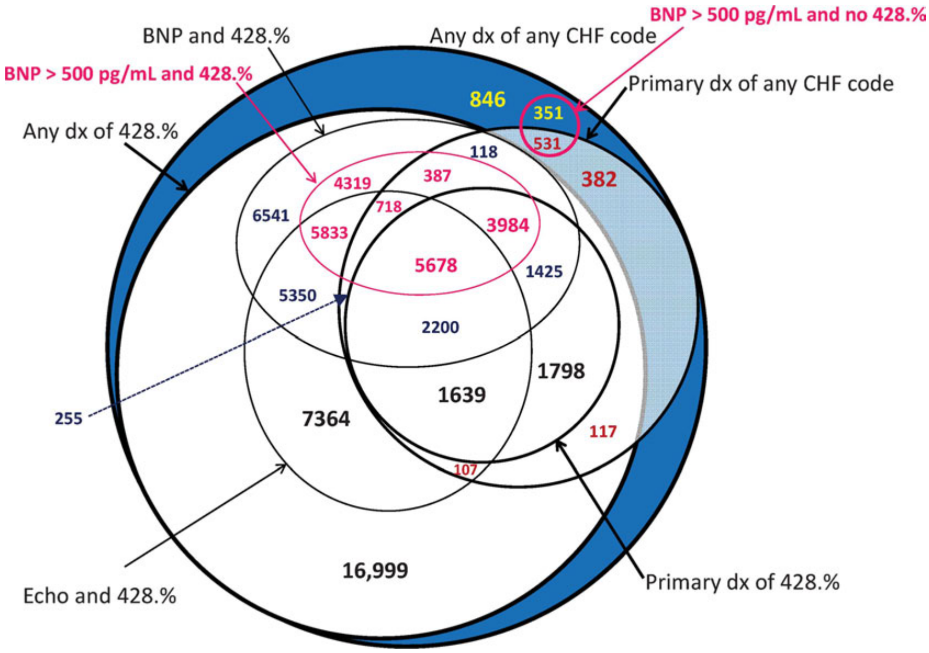

Figure 1 Number of hospitalizations by Venn diagram zone. BNP, B-natriuretic peptide; CHF, congestive heart failure.

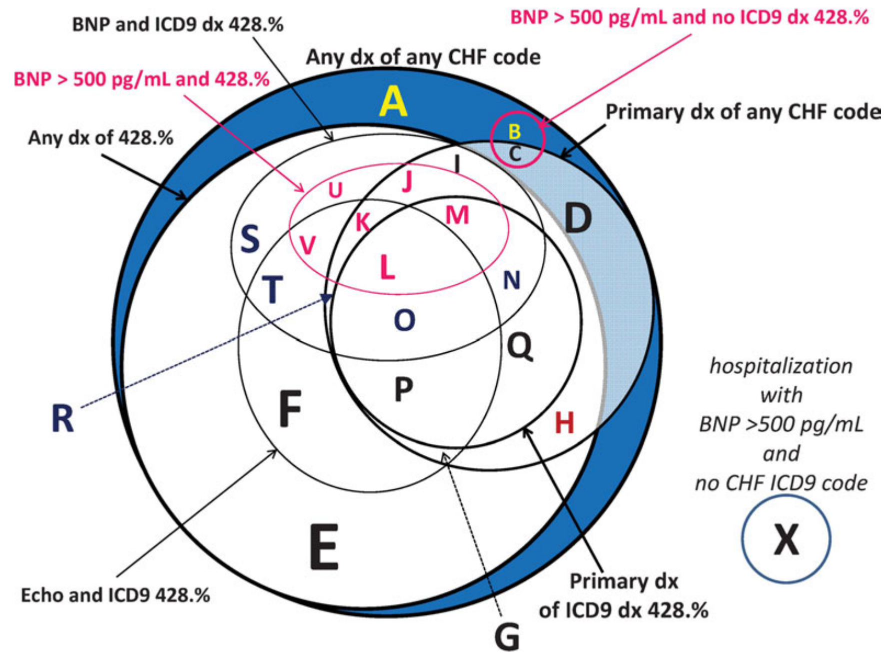
BNP and ICD9 dx 428.%
BNP > 500 pg/mL and no ICD9 dx 428.%
Any dx of any CHF code
BNP > 500 pg/mL and 428.%
A
B
C
Primary dx of any CHF code
Any dx of 428.%
I
U
J
S
V
K
M
D
L
T
N
O
Q
R
F
P
H
hospitalization with BNP >500 pg/mL and no CHF ICD9 code
X
E
Echo and ICD9 428.%
Primary dx of ICD9 dx 428.%
G

Figure 2 Venn diagram zones. BNP, B-natriuretic peptide; CHF, congestive heart failure.

Sampling strategy

For the purpose of validation, our primary goal is to use manual review of charts to estimate the prevalence of true cases (PPV) and the sensitivity (proportion of all true cases identified) in a Venn diagram region (combination of zones) defined by an algorithm. Since available resources generally do not permit manual review of all cases identified by the various algorithms, a random sample must be chosen from the population. To estimate the prevalence of true positives in a given region, simple random sampling in the region results in a larger variance of the estimate compared to the variance that would arise from proportional sampling (using the same proportion across each zone within the region) with the same total sample size.[30] For example, to estimate the measures of accuracy of the algorithm ‘any dx of any CHF ICD-9 code’ (first algorithm in table 1, zones A-through-V in figure 2), proportional sampling of all the zones is more efficient than simple random sampling from the entire large circle.

Table 1 Ten inclusion criteria sets, and corresponding zones of the Venn diagram

Inclusion criteria	Zones
Any dx of any CHF ICD-9 code	(A-through-V)
Any dx of ICD-9 code 428	(E-through-V)
Any dx of any CHF ICD-9 code and any BNP >500 pg/mL	(B, C, J, K, L, M, U, V)
1^0 dx of any CHF ICD-9 code	(C, D, G-through-R)
1^0 dx of ICD-9 code 428	(L-through-Q)
1^0 dx of any CHF ICD-9 code, and BNP >500 pg/mL	(C, J, K, L, M)
1^0 dx of ICD-9 code 428, and BNP >500 pg/mL	(L, M)
1^0 dx of ICD-9 code 428, and BNP >500 pg/mL, and Echo	L
1^0 dx of any CHF ICD-9 code, or BNP >500 pg/mL	(B, C, D, G-through-Q, U, V)
1^0 dx of ICD-9 code 428, or BNP >500 pg/mL	(B, C, J, K, L-through-Q, U, V)

BNP, B-natriuretic peptide; CHF, congestive heart failure.

However, there are situations when we want to compare the PPV or sensitivity of two different algorithms that are overlapping, but one is not a strict subset of the other. For example, we may be interested in comparing the algorithm of ‘primary diagnosis of ICD-9 code 428’ (zones L-through-Q in figure 2, and row 5 in table 1) against the alternative of ‘primary diagnosis of any CHF ICD-9 code and BNP >500 pg/mL’ (zones C, J-through-M, in figure 2, and row 6 in table 1). The latter is less restrictive in ICD-9 codes but more restrictive by the addition of the BNP criterion. Since the overlapping region (zones L, M) contributes to both algorithms, the non-overlapping regions—(zones C, J, and K) and (zones N-through-Q)—are the source of the differences in prevalence and sensitivity. To ensure adequate power to detect such differences, we should sample a sufficient number of cases in all non-overlapping regions between pairs of algorithms that may be compared. In the acute CHF example, comparisons between all possible pairs of algorithms in table 1 could require estimates from these non-overlapping regions: (A,B,C,D), (E,F), (G,H,I), (J,K), (L), (M), (N through Q), (R), (S,T) and (U,V). Therefore we need adequate sample sizes only in these regions.

To optimize the sampling strategies for estimating PPVs and to maximize power to detect differences in PPVs, we therefore recommend a heuristic approach that is a compromise between proportional sampling and over-sampling in the non-overlapping regions to achieve adequate sample size. In the acute CHF example, the total number of cases identified by any set of criteria (any CHF ICD-9 code) is 66 942. We could afford to perform manual reviews of no more than 1000 charts. Therefore we started with an initial plan with a 1% sample stratified on each zone, along with an increase in the sampling proportion to 5% or 10% in the zones that made up small regions so that the total sample size in each region at least approached 100. On the other hand, a 1% sample in a large zone yielded more than 100; manual reviews of that many charts in a single region would provide diminishing returns in terms of precision. Therefore, if resources were limited during the chart review, we could reduce the sample size in zone E.

Chart reviews

In consultation with CHF-specialist cardiologists, we developed a priori criteria and a questionnaire for the chart reviewers to use in assessing whether acute CHF exacerbation was present or absent during the hospital stay. The chart reviewers used CareWeb (the clinician user-interface for the INPC) to review free-text documents including hospital discharge summaries and admission notes. A structured chart abstraction tool was used to collect the data in Microsoft Access. The chart reviewers included two of the authors (MBR, KN), supplemented by a team of one other physician and four senior medical students. The team was trained by the lead author, and cross-validation of selected charts by the lead author was used to ensure consistency; also, MBR and KN reviewed 67% of all charts. The samples were divided such that at least two reviewers worked on each Venn diagram zone.

The criteria for the presence or absence of acute CHF exacerbation included the treating physician's assessment of why the patient had been hospitalized, the Framingham criteria for acute CHF,[31] and/or whether a positive response to intravenous furosemide had been documented. If acute CHF was judged not to have been present, the reviewer noted whether or not chronic CHF was a co-morbid illness, and also noted what the treating physician(s) had documented as the reason for admission.

Statistical analyses

Estimation of PPV

In evaluating the performance of an algorithm for identifying a phenotype, we do not use the measure specificity, which is defined as, among all true negatives, the proportion which the algorithm labels as negative. This is because the overwhelming number of hospitalizations in an EHR system is unrelated to a particular phenotype, acute CHF in our example, so the specificity always approaches 1.0. Instead, the PPV is a meaningful measure of how specific the algorithm is. PPV is defined as, among all cases identified by an algorithm, the proportion which are true positives.

Let the mutually exclusive zones be labeled $i=1, \ldots, I$, the number of cases in zone i be N_i, with sampling proportion p_i, resulting in the number of cases sampled equal to $n_i=N_ip_i$ in zone i. Out of the n_i charts reviewed, let the number of true positives be x_i, which follows a binomial distribution with total n_i and probability π_i (the true proportion of true positive cases in zone i). For any combination of zones i in a set Z as defined by a given algorithm, the overall PPV is estimated by: $(\sum_{i\epsilon Z} x_i/p_i)/(\sum_{i\epsilon Z} N_i)$ with variance $(\sum_{i\epsilon Z} (N_i/p_i)\pi_i(1-\pi_i))/((\sum_{i\epsilon Z} N_i)^2)$, the square root of which gives the SE, in which (x_i/n_i) can be used as an estimate of π_i.

Estimation of sensitivity

An algorithm could have high PPV but could be so restrictive that it misses the identification of many true cases. Therefore we also want to assess the performance of an algorithm by its sensitivity, defined as the proportion of all true cases that can be identified by the algorithm. It is often not feasible to assess sensitivity because we need to know the number of true cases among the vast EHR system outside of the algorithm being considered; a reliable estimate for that number from a random sample requires a huge sample. Validation studies that have reported sensitivities generally have an independent list of known true positives and estimate the sensitivity as the proportion of the list that can be identified by the algorithm. If we had previously performed a clinical study of hospitalized patients evaluated for acute CHF, we would have had a list with which to evaluate the sensitivities of our algorithms. Without such a list, we assumed that any CHF hospitalizations not captured by any of our algorithms had to have had BNP >500 pg/mL (zone X in figure 2). We performed chart reviews on a sample of this group to estimate the total number of true cases that escaped identification by any of our algorithms.

For an algorithm that corresponds to a region Z made up of certain zones in the Venn diagram, the numerator of the sensitivity is the estimated number of true positives in Z, given by $\sum_{i\epsilon Z} x_i/p_i$. The denominator, that is, the estimated number of true positives in the entire EHR system, is common to all algorithms, and is given by $\sum_{i\epsilon Y} x_i/p_i$, where Y denotes all zones from which samples were drawn, including the zones outside of all algorithms (zone X in our example). Thus the sensitivity is given by $\{\{\sum_{i\epsilon Z} x_i/p_i\}/\{\sum_{i\epsilon Y} x_i/p_i\}$. The approximate variance from the delta process is given by: $(N_0^2/(N_o+N_1)^4)\,\mathrm{var}\,(N_1)+(N_1^2/(N_o+N_1)^4)\,\mathrm{var}\,(N_0)$, where the estimated numbers of true positives inside and outside the region of an algorithm are, respectively, $N_0=\sum_{i\in Z} N_ip_i$, $N_1=\sum_{i\in Y\mathrm{not}Z} N_ip_i$, with respective variances $\mathrm{var}\,(N_0)=\sum_{i\in Z} N_i^2p_i(1-p_i)/n_i$, and $\mathrm{var}(N_1)=\sum_{i\in Y\mathrm{not}Z} N_i^2p_i(1-p_i)/n_i$.

Comparisons between algorithms

To test the PPV between any two algorithms, we simply take the difference of the two estimated PPVs. The SE of the difference estimate is given by the square root of the sum of their variances. The equality of the two PPVs can also be tested using the Z test formed by dividing the difference by its SE. The sensitivities of two algorithms can be similarly compared.

RESULTS

At the two clinical institutions studied, we found 66 942 hospitalizations with at least one ICD-9 diagnosis for CHF; we also found 12 149 hospitalizations with no ICD-9 diagnosis for CHF but a maximum BNP level >500 pg/mL. We reviewed charts for 810 of the 66 942, and 98 of the 12 149 hospitalizations. Table 2 shows the number of hospitalizations in each Venn diagram zone, and the number sampled.

The results for the 10 queries of interest are shown in table 3. A query for *any* ICD-9 code for CHF had PPV 42.8% (SE 1.5%) for acute CHF and sensitivity 94.3% (1.3%). A query for a *primary* diagnosis of 428 and BNP >500 pg/mL had PPV 90.4% (SE 2.4%) and sensitivity 28.8% (1.1%).

When we examined other areas of the Venn diagram (particular zones, or combinations of zones not included in the 10 queries of interest above), the PPV varied widely. To cite one example, in zone E (hospitalizations with an ICD-9 diagnosis of 428 which was not primary, and no echocardiogram, and no BNP), the PPV was 8%.

Comparing 'primary diagnosis of ICD-9 428' with 'primary diagnosis of any CHF and BNP >500 pg/mL', the latter algorithm has higher PPV=(90.0–86.3%)=3.7%, with SE of $\sqrt{2.1^2+2.5^2}=3.3\%$. A Z test (with Z-statistic of 3.7/3.3=1.1) of the difference had p value >0.1, so the PPV of these two algorithms did not differ significantly. In contrast, the algorithm 'primary diagnosis of ICD-9 428' has significantly higher sensitivity by 14.1% (47.6%–33.5%) with SE of $\sqrt{1.7^2+1.3^2}=2.1\%$, so it is a better choice than 'primary diagnosis of any CHF and BNP >500 pg/mL'.

Two ellipses in the Venn diagram—(1) echocardiogram and any diagnosis of 428, and (2) BNP level drawn and any diagnosis of 428—were not used, in and of themselves, as phenotype queries (in tables 1 and 3). These two algorithms may be

Table 2 Sampling numbers by Venn diagram zone

Venn zone	N in zone	1% sample	Target sampling proportion (%)	Goal N for sampling	Actual N sampled
A	846	9	5	42	32
B	351	4	5	18	15
C	531	6	5	27	19
D	382	4	5	19	19
E	16 999	170	1	170	134
F	7364	74	1	74	63
G	107	2	5	5	3
H	117	2	5	6	5
I	118	2	5	6	6
J	387	4	10	39	25
K	718	8	10	72	54
L	5678	57	2	114	97
M	3984	40	2	80	47
N	1425	14	1	14	9
O	2200	22	1	22	17
P	1639	16	1	16	17
Q	1798	18	1	18	18
R	255	3	10	26	21
S	6541	65	1	65	59
T	5350	54	1	54	49
U	4319	43	1	43	50
V	5833	58	1	58	51
All combined	66 942	669		988	810
High BNP, no ICD-9 diagnosis for CHF					
X	12 149	N/A	N/A	91	98

ICD-9 diagnosis for congestive heart failure (CHF) present.
BNP, B-natriuretic peptide.

valuable in a database composed exclusively of healthcare payer claims; however, in this HIE with both clinical and payer data, we focused our algorithm selection and sampling considerations on BNP *results* (>500 pg/mL) and used the presence of an echocardiogram as an 'and' statement in one of the algorithms. But since the *existence* of a BNP result or echocardiogram may have value in phenotype algorithms in healthcare payer databases, we conducted a separate analysis of these two ellipses and their intersection (table 4).

On chart review, 'false positive' hospitalizations were most commonly for other heart disease (eg, coronary disease, arrhythmia without acute CHF) or lung disease (eg, exacerbation of chronic obstructive pulmonary disease (COPD), pneumonia). However, as figure 3 shows, there was a wide range of other reasons for admission.

As expected, the chart reviewers encountered hundreds of different ways that physicians had described the presence or absence of acute CHF. Some of the turns of phrase were, from true positive cases, 'frank CHF' or 'profound diuresis' (after furosemide), and, from 'false positive' hospitalizations, 'I doubt he is in active heart failure at the moment'. We recognized various false-positive clinical scenarios. The most salient involved patients with fluid overload because of renal failure. Some patients had missed their scheduled dialysis session. Phrases like 'missed dialysis' or 'dialysis catheter fell out', while not dispositive for the absence of acute CHF, seemed to have a negative predictive value high enough to consider in our next project—one involving the contribution that natural language processing might make to our Venn diagram methods.

DISCUSSION

Developing and validating phenotypes in EHR data is time-consuming. The eMERGE network recently underscored the 'value of iterative algorithm development', along with chart review and calculation of PPV, when validating and 'fine-tuning' a phenotype.[10] We agree. We also submit that our Venn diagram methods can—for at least some clinical phenotypes—increase the efficiency of the work. The principal advantage is that we do not have to commit to a single set of phenotype criteria at the outset. Rather, one set of chart reviews can be used to validate multiple overlapping criteria. Subsequent users of the phenotype will have various purposes and may desire the flexibility conferred by having information about the test characteristics of multiple overlapping options.

As clinical phenotypes take on many important roles in the *Big Data* era ahead,[32] we anticipate that it will be increasingly important to estimate the sensitivity, not only the PPV. Our methods enable derivation of point estimates and SEs of the sensitivity, for the various overlapping queries, based on the assumption that zone X in the Venn diagram contains all the missed true cases. Without a 'zone X', sensitivity may be cost-prohibitive to calculate. In a large EHR or HIE, to reliably estimate the sensitivity by reviewing a *random* sample of all records

Table 3 Results for the 10 congestive heart failure (CHF) phenotype queries

Criteria to combine Venn diagram zones	N in query	Sensitivity (%)	Sensitivity, SE (%)	PPV (%)	PPV, SE (%)
Any CHF	66 942	94.3	1.3	42.8	1.5
Any dx of 428	64 832	90.9	1.3	42.5	1.5
Any dx of CHF and BNP >500 pg/mL	21 801	50.8	1.8	70.7	2.5
1^0 dx of any CHF	19 339	54.8	1.9	86.0	2.2
1^0 dx of 428	16 724	47.6	1.7	86.3	2.5
1^0 dx of any CHF and BNP >500 pg/mL	11 298	33.5	1.3	90.0	2.1
1^0 dx of 428 and BNP >500 pg/mL	9662	28.8	1.1	90.4	2.4
1^0 dx of 428 and BNP >500 pg/mL and echocardiogram	5678	16.2	0.8	86.6	3.5
1^0 dx of any CHF or BNP >500 pg/mL	29 587	71.4	2.1	73.3	2.2
1^0 dx of 428 or BNP >500 pg/mL	28 863	69.6	2.1	73.2	2.2
High BNP, no ICD-9 diagnosis for CHF					
Zone X: no ICD-9 dx of 428, but BNP >500 pg/mL	12 149	N/A	N/A	14.3	3.5

BNP, B-natriuretic peptide; PPV, positive predictive value.

Table 4 Additional results (existence of BNP and/or echocardiogram)

Criteria to combine Venn diagram zones	N in query	Sensitivity (%)	Sensitivity, SE (%)	PPV	PPV, SE (%)
BNP exists and 428 exists	36 808	71.6	1.9	59.0	2.1
Echo exists and 428 exists	29 144	54.6	1.8	56.8	2.5
BNP exists and 428 exists and echo exists	20 034	45.0	1.7	68.2	3.0

BNP, B-natriuretic peptide; PPV, positive predictive value.

not identified by the phenotype query would require a huge sample. But selecting a 'zone X' may be somewhat arbitrary and may be sometimes impossible. In a similar study, we wrote a phenotype query for outpatients with diabetic retinopathy; ICD-9 diagnoses were readily available, but electronic elements for a 'zone X' based on ophthalmology clinic data were not available in the INPC.

In the acute CHF example presented here, the assumption that 'false negatives' would always have a BNP >500 pg/mL is of course a limitation. BNP itself is a parameter with its own predictive relationship with acute CHF and may be affected by body mass index, age, and kidney function.[28 29 33 34] Because some patients with BNP levels in the 100–500 pg/mL range (and no CHF ICD-9 diagnosis) may have had acute CHF, our reported sensitivities may be overestimates. However, only 14% of the hospitalizations in zone X had acute CHF, even with the strict >500 pg/mL threshold. Lowering the threshold to 100 pg/mL would increase the number of hospitalizations in zone X but would further lower the percentage there with acute CHF.

The validity of ICD-9 diagnosis-based queries for CHF varies in previous studies. A review by Quach *et al*,[35] of 25 administrative data studies, found a wide range of PPV and sensitivities. Goff *et al*,[36] in a study of 5100 cardiac-related hospitalizations (for myocardial infarction, coronary artery disease, dysrhythmias, chest pain) from 1988–1994, found that CHF ICD-9 diagnoses had PPV 77% and sensitivity 67% for acute CHF. The higher PPV and lower sensitivity than ours probably reflect differences in the methods. The higher PPV makes sense in that Goff's overall pool of hospitalizations was limited to cardiac-related hospitalizations. The lower sensitivity may reflect other etiologies of pulmonary edema (eg, renal) within the gold standard. The 'zone X' equivalent in the Goff study—hospitalizations with no CHF ICD-9 code but any of a variety of other heart-related ICD-9 codes—may have been more expansive than ours. The Goff study predated the BNP era.[37–40]

More recently, Li *et al*,[41] using linked Medicare and CHF/myocardial infarction registry data, studied 13 queries with a narrower range of CHF ICD-9 code definitions (and some with diagnosis-*and*-medication criteria), with systolic dysfunction as 'true positive' (ejection fraction <45%). Not unexpectedly based on these definitions, their PPVs were generally higher, and sensitivities were lower, than ours. Our results based on ICD-9 codes alone are perhaps closest to those of Schellenbaum *et al*, who, in a study of incident CHF, compared CHF ICD-9 diagnoses against adjudications made by a five-physician events committee; among 1072 hospitalizations with a CHF ICD-9 diagnosis, the five-physician events committee validated 575 (54%) as CHF.[42]

A query for acute CHF could be modified by applying additional criteria. The chart reviews showed that some of the hospitalizations were for renal failure or for COPD. We reran the automated queries to explore whether excluding hospitalizations with an ICD-9 diagnosis for renal failure (584.xx, 585.xx, or 586.xx) or COPD (492.xx or 496.xx) might improve the PPVs of the queries for acute CHF. The result was that the PPVs for acute CHF were little changed. The PPV of the 'any CHF' query was 39.8% when renal failure ICD-9 codes were excluded and 43.0% when COPD ICD-9 codes were excluded (versus 42.8% without these exclusions). As with the ICD-9 codes for CHF, the ICD-9 codes for renal failure and COPD are imprecise; more sophisticated queries for these exclusion phenotypes will be required. The Venn region where excluding renal failure ICD-9 codes made the biggest difference (an increase in PPV from 87.2% to 91.4%) was in (zones J, K)—that is, when there was a high BNP and a primary diagnosis of one of the non-428.xx CHF ICD-9 codes. That result makes sense, in that the non-428.xx CHF ICD-9 codes include 404.x1 and 404.x3 (heart failure with hypertensive heart and chronic kidney disease).

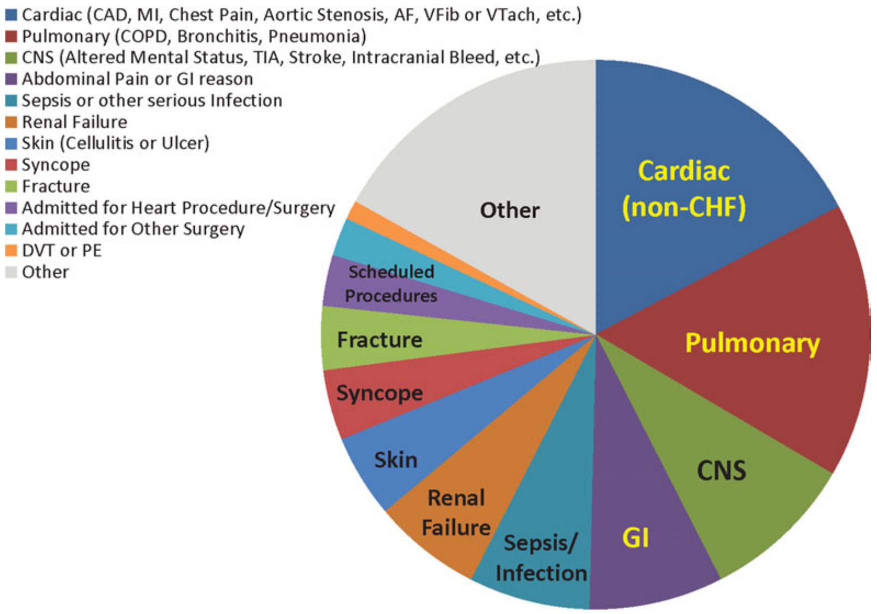

Figure 3 Reasons for admission when it was not acute congestive heart failure (CHF).

Another query criterion one could explore is to specify the hospital (or hospital system). We observed what appears to be a substantial difference in the PPVs across the two clinical institutions in this study. (When the phenotype was based on a query for primary diagnosis of ICD-9 code 428, the absolute difference in PPV between the two institutions was 12.5%. When the phenotype was based on a query for primary diagnosis of ICD-9 code 428 and BNP >500 pg/mL, the absolute difference in PPV between the two institutions was 8.6%.) This finding reinforces our experience that phenotype queries vary in performance across settings, based on differences in the quantity or quality of data, the patient mix, the clinical workflows, or other factors. Several studies from the eMERGE network illustrate the transportability of phenotype algorithms across the (heterogeneous) eMERGE centers and the opportunities that having multi-source data affords for retraining and refining the algorithms.[10 11 43–45] A multi-institution regional HIE—even one committed to standards for data and messaging—may encompass even more heterogeneity than a network like eMERGE. The Venn diagram and set theory methods may be useful in studying the transportability of electronic phenotype queries and in elucidating variation across institutions or cities.

Our approach has some additional limitations or nuances to mention. There is not necessarily one clear choice of a 'gold standard' for validation. In addition to the Framingham criteria,[31] there are others that we could have selected.[46–50] As seen in table 1, we did not quite always reach our chart review sample goal in each zone; we had a short project timeline, and assignments were divided among multiple team members. Still, we adapted our assignments along the way to ensure reasonable and balanced samples. Sometimes also, a detailed chart review in a Venn diagram framework helps suggest ways that the preceding, automated phenotype queries could be improved. This opportunity helped us revise and improve our queries in this project.

We agree with Hripcsak and Albers that the 'full challenge of phenotyping is not broadly recognized'.[9] In this project, we reduced some complexity by using only two clinical institutions for the hospitalizations in the INPC. The Venn diagram method may have more value in future studies exploring a wider array of HIE institutions. The first time we used this method, to identify infantile hypertrophic pyloric stenosis,[51] it was only by comparing the non-overlap between ICD-9 code 750.5 (the appropriate code for infants) and a non-ICD-9 circle that we discovered that a hospital was coding some infants not with 750.5 but with a code for adults, 537.0 (acquired hypertrophic pyloric stenosis). In the earliest stages of building a phenotype, the Venn diagram can help reveal the 'physics'[9]—or, in our conception of it, the anthropology—of the EHR/HIE.

Acknowledgements We gratefully acknowledge Layla Baker, MBChB, MPH, CCRP, for her assistance with the medical record reviews.

Contributors MR conceived and designed the study, analyzed and interpreted data, drafted the article and revised it critically for important intellectual content, and approved the final version. JH helped conceive and design the study, revised the article critically for important intellectual content, and approved the final version. JM conducted extensive electronic queries, analyzed and interpreted data, helped revise the article for important intellectual content, and approved the final version. KN conducted extensive chart reviews with abstraction and interpretation of data, helped revise the article for important intellectual content, and approved the final version. GE and KL helped conceive, design, and apply the statistical methods, revised the article critically for important intellectual content, and approved the final version. IG-P helped conceive and design the study, revised the article critically for important intellectual content, and approved the final version. SLH helped conceive the study, designed the statistical methods, analyzed and interpreted data, revised the article critically for important intellectual content, and approved the final version.

Funding This study was funded by Merck Sharp & Dohme.

Competing interests This study was funded by Merck Sharp & Dohme. JH is a scientist employed by Merck Sharp & Dohme.

Ethics approval Indiana University Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Rice JP, Saccone NL, Rasmussen E. Definition of the phenotype. *Adv Genet* 2001:42:69–76.
2 Wojczynski MK, Tiwari HK. Defintion of phenotype. *Adv Genet* 2008;60:75–105.
3 Cutrona SL, Toh S, Iyer, *et al*. Validation of acute myocardial infarction in the Food and Drug Administration's Mini-Sentinel program. *Pharmacoepidemiol Drug Saf* 2013;22:40–54.
4 Overhage JM, Ryan PB, Reich CG, *et al*. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc* 2012;19:54–60.
5 Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann Intern Med* 2010;153:600–6.
6 National Quality Forum. Electronic Quality Measures (eMeasures). http://www.qualityforum.org/Projects/e-g/eMeasures/Electronic_Quality_Measures.aspx. (accessed Mar 2013).
7 National Committee for Quality Assurance. HEDIS and performance measurement. http://www.ncqa.org/HEDISQualityMeasurement.aspx (accessed Mar 2013).
8 Kern LM, Malhotra S, Barrón Y, *et al*. Accuracy of electronically reported 'meaningful use' clinical quality measures. A cross-sectional study. *Ann Intern Med* 2013;158:77–83.
9 Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc* 2013;20:117–21.
10 Newton KM, Peissig PL, Kho AN, *et al*. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013;20:e147–54.
11 Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19:e162–9.
12 Kahn MG, Raebel MA, Glanz JM, *et al*. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. *Med Care* 2012;50(Suppl):S21–9.
13 Desai JR, Wu P, Nichols GA, *et al*. 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:S30–5.
14 Jollis JG, Ancukiewicz M, DeLong ER, *et al*. Discordance of databases designed for claims payment versus clinical information systems. Implications for outcomes research. *Ann Intern Med* 1993;119:844–50.
15 Tang PC, Ralston M, Arrigotti MF, *et al*. 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;14:10–15.
16 Jones N, Schneider G, Kachroo S, *et al*. A systematic review of validated methods for identifying acute respiratory failure using administrative and claims data. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):261–4.
17 Schneider G, Kachroo S, Jones N, *et al*. A systematic review of validated methods for identifying hypersensitivity reactions other than anaphylaxis (fever, rash, and lymphadenopathy), using administrative and claims data. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):248–55.
18 Cutrona SL, Toh S, Iyer A, *et al*. Design for validation of acute myocardial infarction cases in Mini-Sentinel. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):274–81.
19 Curtis JR, Mudano AS, Solomon DH, *et al*. Identification and validation of vertebral compression fractures using administrative claims data. *Med Care* 2009;47:69–72.
20 McDonald CJ, Overhage JM, Barnes M, *et al*. The Indiana Network for Patient Care: a working local health information infrastructure. *Health Affairs* 2005;24:1214–20.
21 Render ML, Almenoff PL, Christianson A, *et al*. A hybrid Centers for Medicaid and Medicare service mortality model in 3 diagnoses. *Med Care* 2012;50:520–6.
22 Joynt KE, Harris Y, Orav EJ, *et al*. Quality of care and patient outcomes in critical access rural hospitals. *JAMA* 2011;306:45–52.
23 Muehlenbein CE, Hoverman JR, Gruschkus SK, *et al*. Evaluation of the reliability of electronic medical record data in identifying comorbid conditions among patients with advanced non-small cell lung cancer. *J Cancer Epidemiol* 2011;2011:983271.
24 Kottke TE, Baechler CJ. An algorithm that identifies coronary and heart failure events in the electronic health record. *Prev Chronic Dis* 2013;10:E29.
25 Amarasingham R, Moore BJ, Tabak YP, *et al*. An automated model to identify heart failure patients at risk for 30-day readmission or death using electronic medical record data. *Med Care* 2010;48:981–8.
26 Keyser DJ, Dembosky JW, Kmetik K, *et al*. Using health information technology-related performance measures and tools to improve chronic care. *Jt Comm J Qual Patient Saf* 2009;35:248–55.

27 American College of Cardiology Foundation (ACCF) American Heart Association (AHA) Physician Consortium for Performance Improvement (PCPI™). Heart Failure Performance Measurement Set. Updated 15 May 2012. http://www.ama-assn.org/ama1/pub/upload/mm/pcpi/hfset-12-5.pdf (accessed Apr 2013).

28 Maisel AS, Krishnaswamy P, Nowak RM, *et al*. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. *N Engl J Med* 2002;347:161–7.

29 Hainaut C, Gade W. The emerging roles of BNP and accelerated cardiac protocols in emergency laboratory medicine. *Clin Lab Sci* 2003;16:166–79.

30 Scheaffer RL, Mendenhall W, Ott RL, *et al*. *Elementary survey sampling*. 7th edn. Duxbury Resource Center, 2011.

31 McKee PA, Castelli WP, McNamara PM, *et al*. The natural history of congestive heart failure: the Framingham study. *N Engl J Med* 1971;285:1441–6.

32 Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.

33 Ray P, Delerme S, Jourdain P, *et al*. Differential diagnosis of acute dyspnea: the value of B natriuretic peptides in the emergency department. *QJM* 2008;101:831–43.

34 Januzzi JL Jr, Maisel AS, Silver M, *et al*. Natriuretic peptide testing for predicting adverse events following heart failure hospitalization. *Congest Heart Fail* 2012;18: S9–S13.

35 Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. *Can J Cardiol* 2010;26:306–12.

36 Goff DC, Pandey DK, Chan FA, *et al*. Congestive heart failure in the United States: Is there more than meets the I(CD code)? The Corpus Christi Heart Project. *Arch Intern Med* 2000;160:197–202.

37 Valli N, Gobinet A, Bordenave L. Review of 10 years of the clinical use of brain natriuretic peptide in cardiology. *J Lab Clin Med* 1999;134:437–44.

38 Mukoyama M, Nakao K, Saito Y, *et al*. Increased human brain natriuretic peptide in congestive heart failure. *N Engl J Med* 1990;323:757–8.

39 Arad M, Elazar E, Shotan A, *et al*. Brain and atrial natriuretic peptides in patients with ischemic heart disease with and without heart failure. *Cardiology* 1996;87: 12–7.

40 Grantham JA, Burnett JC. BNP: increasing importance in the pathophysiology and diagnosis of congestive heart failure. *Circulation* 1997;96:388–90.

41 Li Q, Glynn RJ, Dreyer NA, *et al*. Validity of claims-based definitions of left ventricular systolic dysfunction in Medicare patients. *Pharmacoepidemiol Drug Saf* 2011;20:700–8.

42 Schellenbaum GD, Heckbert SR, Smith NL, *et al*. Congestive heart failure incidence and prognosis: case identification using central adjudication versus hospital discharge diagnoses. *Ann Epidemiol* 2006;16:115–22.

43 Kho AN, Hayes MG, Rasmussen-Torvik L, *et al*. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc* 2012;19:212–8.

44 Peissig PL, Rasmussen LV, Berg RL, *et al*. Importance of multi-modal approaches to effectively identify cataract cases from electronic health records. *J Am Med Inform Assoc* 2012;19:225–34.

45 Denny JC, Crawford DC, Ritchie MD, *et al*. Variants near FOXE1 are associated with hypothyroidism and other thyroid conditions: using electronic medical records for genome- and phenome-wide studies. *Am J Hum Genet* 2011;89:529–42.

46 Marantz PR, Tobin JN, Wassertheil-Smoller S, *et al*. The relationship between left ventricular systolic function and congestive heart failure diagnosed by clinical criteria. *Circulation* 1988;77:607–12.

47 Harlan WR, Oberman A, Grimm R, *et al*. Chronic congestive heart failure in coronary artery disease: clinical criteria. *Ann Intern Med* 1977;86:133–8.

48 Carlson KJ, Lee DC, Goroll AH, *et al*. An analysis of physicians' reasons for prescribing long-term digitalis therapy in outpatients. *J Chronic Dis* 1985;38:733–9.

49 Schellenbaum GD, Rea TD, Heckbert SR, *et al*. Survival associated with two sets of diagnostic criteria for congestive heart failure. *Am J Epidemiol* 2004;160:628–35.

50 Ives DG, Fitzpatrick AL, Bild DE, *et al*. Surveillance and ascertainment of cardiovascular events: the Cardiovascular Health Study. *Ann Epidemiol* 1995;5:278–85.

51 Mahon BE, Rosenman MB, Kleiman MB. Maternal and infant use of erythromycin and other macrolide antibiotics as risk factors for infantile hypertrophic pyloric stenosis. *J Pediatr* 2001;139:380–4.