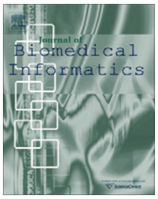
Journal of Biomedical Informatics 43 (2010) 914–923

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



An analytical approach to characterize morbidity profile dissimilarity between distinct cohorts using electronic medical records

Jonathan S. Schildcrout [a,b,*], Melissa A. Basford [c], Jill M. Pulley [d], Daniel R. Masys [d,e], Dan M. Roden [d,f], Deede Wang [d], Christopher G. Chute [g], Iftikhar J. Kullo [h], David Carrell [i], Peggy Peissig [j], Abel Kho [k], Joshua C. Denny [d,e]

[a] Department of Biostatistics, Vanderbilt University School of Medicine, United States
[b] Department of Anesthesiology, Vanderbilt University School of Medicine, United States
[c] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University School of Medicine, United States
[d] Department of Medicine, Vanderbilt University School of Medicine, United States
[e] Department of Biomedical Informatics, Vanderbilt University School of Medicine, United States
[f] Department of Pharmacology, Vanderbilt University School of Medicine, United States
[g] Division of Biostatistics and Informatics, Mayo Clinic, United States
[h] Division of Cardiovascular Diseases, Mayo Clinic, United States
[i] Center for Health Studies, Group Health Cooperative, United States
[j] Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, United States
[k] Department of Internal Medicine, Northwestern University School of Medicine, United States

ARTICLE INFO

Article history:
Received 6 November 2009
Available online 3 August 2010

Keywords:
Electronic medical records
ICD-9
Dissimilarity index
Comorbidity index
Population comparison
Morbidity dissimilarity index

ABSTRACT

We describe a two-stage analytical approach for characterizing morbidity profile dissimilarity among patient cohorts using electronic medical records. We capture morbidities using the International Statistical Classification of Diseases and Related Health Problems (ICD-9) codes. In the first stage of the approach separate logistic regression analyses for ICD-9 sections (e.g., "hypertensive disease" or "appendicitis") are conducted, and the odds ratios that describe adjusted differences in prevalence between two cohorts are displayed graphically. In the second stage, the results from ICD-9 section analyses are combined into a general morbidity dissimilarity index (MDI). For illustration, we examine nine cohorts of patients representing six phenotypes (or controls) derived from five institutions, each a participant in the electronic MEdical REcords and GEnomics (eMERGE) network. The phenotypes studied include type II diabetes and type II diabetes controls, peripheral arterial disease and peripheral arterial disease controls, normal cardiac conduction as measured by electrocardiography, and senile cataracts.

1. Introduction

Electronic medical records (EMR) have been shown to offer the potential to improve the quality of clinical care, reduce costs, and improve guideline adherence. While researchers have also used EMRs for clinical research [1,2], for medical outcomes research [3], to categorize rare findings [4], and to identify patients with various conditions and assess eligibility for clinical trials [5,6], there has been little exploration of using DNA biobanks linked to EMRs for genomic studies. Given the powerful potential for substantial cost and time efficiency [7], there is increasing interest in EMRs as a potential way to identify cohorts of patients and associated DNA samples to discover genetic associations for common complex diseases and the genetic influence on response to therapy through genome-wide association studies (GWAS) [8].

Pooling data from multiple EMRs or sites can improve power and generalizability, especially when investigating a less prevalent disease phenotype. However, it introduces analytical considerations related to cohort heterogeneity. If genotype-phenotype associations are highly variable across the sites, caution should be applied when combining results since a single summary measure of the overall association may mask important site-by-genotype interactions. When a single association measure is of interest, meta-analytic approaches such as the random effects model of DerSimonian and Laird [9] and its extensions can be applied. In this model, the overall association (e.g., a log odds ratio), θ, is a weighted average of the site-specific associations, θ_i, where $i = 1,2,\ldots,I$ denotes site. The variance of θ, $Var(\theta)$, is given by $\left(\sum_{i=1}^{I} 1/(\sigma_i^2+\tau^2)\right)^{-1}$ where $\sigma_i^2 \equiv Var_i(\theta_i)$ is the variance at site i

[*] Corresponding author. Address: 1161 21[st] Ave South, S-2323 Medical Center North, Vanderbilt University School of Medicine, Nashville, TN 37232-2156, United states. Fax: +615 343 4924.
E-mail address: jonathan.schildcrout@vanderbilt.edu (J.S. Schildcrout).

1532-0464/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2010.07.011

and τ^2 is a measure of variability among θ_i across the sites. The value τ^2 can be thought of as a heterogeneity penalty that increases $Var(\theta)$ and can lead to diminished power to detect associations. If costs associated with ascertaining genotypes and/or phenotypes are high, being able to anticipate analytical challenges and/or loss of power due to cohort heterogeneity is crucial. Towards that end, we propose a two-stage analysis protocol that uses readily available patient information to proactively examine the extent to which selected cohorts are dissimilar over a (broad or narrow) range of morbidities.

Due to their wide availability, standard format, and relatively consistent utilization, we capture morbidities with the International Statistical Classification of Diseases and Related Health Problems codes (ICD-9). However, the proposed approach is general and can be applied to other morbidity definitions. At the first stage, the protocol estimates demographic adjusted measures of cohort morbidity differences across individual ICD-9 sections using logistic regression and displays odds ratios and associated 95% confidence intervals graphically. At the second stage, the section-specific differences estimated at the first stage are combined into a single, general measure of cohort dissimilarity. We call this the "morbidity dissimilarity index" (MDI), and it can be thought of as a distance between the morbidity profiles of two cohorts. Results from the two stages of analyses are complementary. Stage 2 results permit broad summarization of dissimilarity over a range of morbidities, and stage 1 results can be used to examine observed differences at a finer level.

2. Background

2.1. Comorbidity summarization

Comorbidity information readily available in EMRs can be a valuable resource for assessing cohort dissimilarity. Individual level indices that can be derived from EMR such as the Charlson comorbidity index [10], Elixhauser index [11,12], APACHE score [13], and functional comorbidity index [14] capture health outcomes related risk for a given a set of features. While these measures can be used to compare individuals' risks, they do not specifically measure similarity. For example two individuals with equal risk scores may differ on the items that comprise the score. An information theoretic scoring approach has been proposed [15] for measuring individual case similarities based on patient-specific features. From this, one could calculate a measure of cohort similarity with, say, an intra-class correlation coefficient that captures the relative contributions of between- and within-cohort variation in the scores. However, by first calculating patient-level scores and then summarizing the distribution of these scores, we lose all information about the relatedness or correlation among the components of the score. As we will show, proper acknowledgment of morbidity correlations is crucial for capturing cohort morbidity similarity. Principal Components Analysis (PCA) is commonly used to identify population (genetic) structure [16–19] and can therefore be used to capture cohort morbidity profile dissimilarity like we do. That is, one could use PCA to reduce the dimensionality of the morbidity profile into, say, a single principal component. A distance metric between the cohorts could then be derived from the morbidity-specific coefficients. However, the morbidity-specific coefficients have conditional interpretations, and therefore in the presence of correlated morbidities marginal differences in prevalence between cohorts will be masked. In our two-stage approach, the marginal differences are of interest and are captured and examined explicitly. They are then combined into a single measure of dissimilarity while properly accounting for morbidity correlations.

2.2. Electronic MEdical Records and GEnomics (eMERGE) Network

This work is motivated by ongoing GWAS studies performed as part of the electronic MEdical Records and GEnomics (eMERGE) network, which seeks to use EMR-linked DNA biobanks as their source of cases and controls. The eMERGE network is a consortium of five medical centers, Group Health Cooperative (GHC, Seattle WA), Marshfield Clinic (MAR, Marshfield, WI), Mayo Clinic (MAY, Rochester, MN), Northwestern University (NU, Chicago, IL), and Vanderbilt University (VU, Nashville, TN). Each eMERGE member has established a DNA biobank linked to an EMR for clinical data [20]. The consortium is funded by the National Human Genome Research Institute with additional funding by the National Institute of General Medical Sciences to develop the necessary tools and techniques to perform GWAS in participants with phenotypes and environmental exposures derived from EMRs.

The eMERGE sites are investigating seven primary disease phenotypes by GWAS, and a growing number of secondary phenotypes that seek to reuse GWAS data derived from the primary phenotypes. Each site has created and refined electronic phenotype selection algorithms to identify cases and controls using information derived from the EMR. The algorithms use combinations of administrative billing codes, laboratory and medication data, and string queries and natural language processing techniques applied to unstructured, free-text clinical narratives. Given the typically small effect size of individual SNP-phenotype associations, thousands of cases and controls are typically required to ensure adequate statistical power for successful GWAS [21]. Thus, several eMERGE phenotypes require pooling cases and controls across the network.

3. Methods

3.1. Populations examined

Across the eMERGE network, selection algorithms were developed for type 2 diabetes (VU, NU), cardiac conduction (VU, NU), senile cataracts (MAR, GHC), senile dementia (MAR, GHC), and peripheral arterial disease (MAY). Each phenotype selection algorithm was iteratively developed and evaluated by clinician reviewers or chart abstractors at each site until they performed well enough to obtain a positive predictive value greater than or equal to 95%. The details of these algorithms are posted on http://gwas.net; their implementation and rationale will be presented in subsequent publications. Because EMR systems and structures differ across sites within the eMERGE network, the algorithms implemented at multiple sites were adapted to accommodate each local environment.

As an example for our analysis protocol, we examined nine site-phenotype cohorts defined by these algorithms: VU type II diabetes (VU-T2D), VU type II diabetes controls (VU-CON), VU patients with normal cardiac conduction as measured by the QRS duration (VU-QRS), NU type II diabetes (NU-T2D), NU type II diabetes controls (NU-CON), GHC senile cataracts (GHC-CAT), MAR senile cataracts (MAR-CAT), MAY peripheral arterial disease (MAY-PAD), and MAY peripheral arterial disease controls (MAY-CON).

3.2. Selection of ICD-9 billing codes for analysis

While billing codes are imperfect measures of disease status, they are useful for research involving EMR because they cover the broad range of diseases and diagnoses, they are commonly used in large scale research to define populations, they are utilized consistently across sites, and they are easily extracted from most EMR systems. Current Procedural Technology (CPT) or ICD-9 procedural codes were not considered because they are dependent on the procedure being performed at the hospital of interest, and the receipt of a procedure is influenced by external factors (e.g., insurance, patient preference, and life expectancy), making them less useful in understanding disease status for many phenotypes. NLP approaches were not applied because these capabilities were not available to all sites in the eMERGE network.

All available inpatient and outpatient ICD-9 codes were selected for each subject and compared against a list of available ICD-9 codes derived from the Unified Medical Language System (UMLS), version 2009AA [22]. Invalid ICD-9 codes, E codes (external causes of injury) V codes (screening codes and other supplementary factors influencing health), procedure codes (i.e., 2-digit ICD9 codes), and signs and symptoms (780–799) were excluded from analyses.

3.3. Data preparation

Adequate EMR data were available for differing lengths of time across eMERGE network sites. For consistency of comparison, the study was limited to the years 2001 to 2007. Five-digit ICD-9 codes were available on all patients, however, coding at this level is highly idiosyncratic, thereby precluding meaningful comparative analyses of the cohorts. On the other hand, regression analyses on codes aggregated to the level of ICD-9 chapters (e.g., "Diseases of the digestive system", $n = 16$) yield coarse and insensitive characterizations of patient co-morbidity profiles. Therefore, to identify co-morbidities, we use ICD-9 categories (3-digit codes, $n = 904$) which we believe represent a level of coding that avoids the major pitfalls of five-digit codes while maintaining sufficient detail to allow meaningful comparisons. For a category code to be considered present in an individual, it must have been observed on more than one occasion. Our rationale for this cut off was (1) it favors chronic conditions over temporary acute conditions, and (2) it reduces potential for noise induced by singular coding errors, as has been found for some chronic conditions in prior ICD-9 analyses [23,24]. While some real co-morbidities might be missed, the approach provides more confidence that the ones observed were indeed true positives. Section (e.g., "Noninfectious enteritis and colitis", $n = 110$) and chapter level co-morbidities were considered to be present if at least one category code underneath them in the ICD-9 taxonomy was present. We only considered adult patients (age $\geqslant$18 years) who were observed for at least three years.

3.4. Analysis strategy

Analyses of ICD-9 categories were considered; however, we found that many important ones did not provide sufficient counts to permit analyses. We base analyses on the 66 of 110 ICD-9 sections that were observed in five percent of patients in at least one cohort and in one-tenth of a percent of patients in all cohorts. Had we not imposed the 'observed category codes twice' rule, our analyses would have been based on 74 ICD-9 sections.

Our analysis protocol involves two stages. In the first stage, we use logistic regression to capture the adjusted log odds ratio of observing each ICD-9 section between cohorts, and in the second stage we summarize section-specific results within and across ICD-9 chapters to ascertain chapter-specific measures and a single overall measure of cohort dissimilarity.

Stage 1: For each ICD-9 section s in 1, 2, ... S, ($S = 66$ in this analysis), we fit a logistic regression model that included, as predictors, the cohort identification variable (i.e., MAY-PAD, NU-T2D, etc.) and covariates: gender, race (white, black, other, and unknown), age, and length of patient follow-up. The demographic covariate adjustments were crucial since multi-site studies include these covariates in their statistical analysis models, and our objective is to characterize cohort morbidity dissimilarity beyond what common adjustment covariates could explain. To reduce re-identification risk, birthdays were truncated to the birth years, and birth years were truncated at 1928. For the sake of modeling, age was represented with two variables: an indicator variable for being born prior to 1928 and then a continuous age variable for those born in or after 1928. The latter age variable and the length of follow-up variable were fit with flexible restricted cubic spline functions with six degrees of freedom [25]. Linear combinations of estimated regression parameters and variances were used to capture differences in the log odds of ICD-9 sections between cohort pairs (e.g., GHC-CAT and MAR-CAT), and the associated odds ratios and confidence intervals were displayed graphically. Because ICD-9 sections were modeled individually, the covariance matrix required for stage 2 was estimated using a stratified bootstrap approach [26]. Specifically, at each of 1500 replicates, a bootstrap sample was ascertained for each site separately, section-specific models were fitted, and parameter estimates were saved for all section-specific models. The covariance matrix was estimated across bootstrap replications [26].

Stage 2: In stage 2, ICD-9 section-specific parameter and covariance estimates from Stage 1 were combined to obtain a measure of cohort dissimilarity. The measure can be described as a rescaled Mahalanobis Distance. Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_S)^t$ be the vector of estimated section-specific differences in the log odds (i.e., the log odds ratio) for two populations estimated at stage 1, and $\hat{V} \equiv \hat{V}(\hat{\beta})$ be the estimated variance–covariance matrix. For ease of exposition, we remove ^ from our notation. We define the morbidity dissimilarity index (MDI) with,

$$\mathrm{MDI} = \sqrt{\beta^t W \beta} = \sqrt{\sum_{s=1}^{S} \sum_{r=1}^{S} \beta_r \beta_s W_{r,s}}$$

where, $W = kV^{-1}$, V^{-1} is the inverse of the variance–covariance matrix V and $k = 1/tr(V^{-1})$ is the inverse of the trace (sum of the diagonal elements) of V^{-1}. The MDI differs from a Mahalonobis distance by the coefficient k, which serves to rescale the measure so it is independent of the magnitude of section-specific variances, and therefore of sample size. Since estimated variances decrease with sample size, Mahalanobis distance necessarily increases with sample size. So, if the goal is classification, the Mahalanobis distance is appropriate; however, our interest is in a simple and interpretable measure of cohort dissimilarity.

When all variances are equal and in the absence of correlation among parameter estimates, the MDI is equal to the Euclidean distance between $(\beta_1, \beta_2, \ldots, \beta_S)$ and the origin (0, 0, ... 0) divided by the square root of S. The MDI is on the same scale as the components of β and therefore, its value has a meaningful interpretation. In contrast, to interpret the Euclidean distance we must know the dimension of β. For example, consider the scenario where $S = 10$ and $\beta = (1, 1, \ldots, 1)$. It is easy to show that the MDI is equal to 1 thereby providing an insightful measure of how large components of β are; however, the Euclidean distance is approximately 3.2, which we find to be less useful.

In the presence of unequal variances and correlation, MDI interpretation is subtle; however proper acknowledgment of these important data features is crucial for characterizing cohort dissimilarity validly. For simplicity, assume we wish to calculate the MDI from analysis of two ICD-9 sections, where σ_1^2 and σ_2^2 are variances for β_1 and β_2 respectively, and ρ is the estimated correlation. It is straightforward to show that the MDI is equal to

$$\mathrm{MDI} = \sqrt{\frac{1}{\sigma_1^2 + \sigma_2^2}(\beta_1^2\sigma_2^2 + \beta_2^2\sigma_1^2 - 2\rho\beta_1\beta_2\sigma_1\sigma_2)}.$$

Upon inspection, it can be seen that MDI does not depend on the magnitude of σ_1^2 and σ_2^2 (i.e., it does not depend on sample size), but it is affected by their relative size and by ρ. Fig. 1 displays the impact of these data features on the MDI, and just as important,

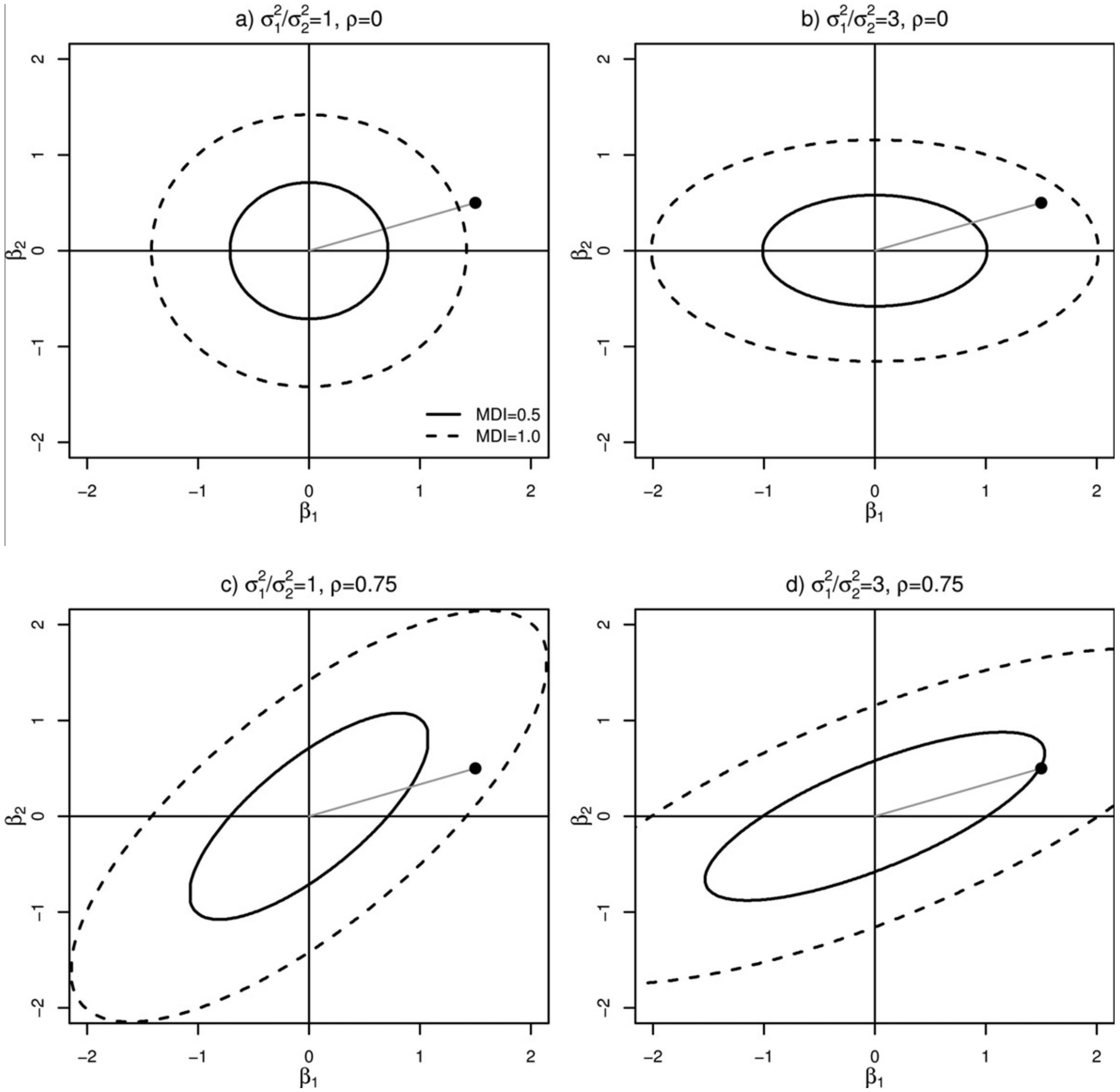


Fig. 1. Example Morbidity Dissimilarity Indices (MDI) for four configurations. MDIs were drawn for $(\sigma_1^2/\sigma_2^2, \rho)$ equal to (1, 0), (3, 0), (1, 0.75), and (3, 0.75) in panels (a), (b), (c), and (d), respectively, where $\sigma_1^2 = Var(\beta_1)$ and $\sigma_2^2 = Var(\beta_2)$ and ρ is the correlation between β_1 and β_2. The solid and dashed contours display the set of all (β_1, β_2), that yield MDI equal to 0.5 and 1.0, respectively.

it shows how misleading dissimilarity measures can be if data features are ignored. Panels are defined by σ_1^2/σ_2^2 and ρ, and in each panel, the solid and dashed black lines display the set of all (β_1, β_2) that result in MDI values equal to 0.5 and 1.0, respectively. Notice that unequal variances stretches or contracts and correlation rotates the parameter space, in that the set of all points corresponding to, say, MDI = 0.5 differs across panels in the figure. The point (1.5, 0.5) is denoted on all panels as a reference point, and the MDI for (1.5, 0.5) in panels (a), (b), (c), and (d), is 1.12, 0.83, 0.86, and 0.51, respectively. That is, if the data structure is given by panel (d), and we ignore the correlation and the differences in variances (e.g., by assuming panel (a) is true) then we will overestimate dissimilarity by more than twofold on the log odds ratio scale. With proper analyses, the MDI effectively addresses unequal variances and correlation. Thus, simpler indices that ignore their impact are not recommended.

4. Results

Demographic characteristics and subject experiences of 17 070 patients observed from January 1, 2001 to December 31, 2007 from eMERGE network sites are shown in Table 1. The NU-T2D cohort was the most racially diverse with minorities representing 36 percent of its sample. The proportion of female subjects ranged from 36% in MAY-PAD to 70% in VU-QRS samples. The GHC-CAT sample was the oldest, with 76 percent of patients being born prior to 1928. This was due to the requirement that patients included in this sample must also qualify for a study on dementia in the elderly. MAY-PAD and MAY-CON cohorts were observed on the fewest number of days with median values equaling 44, while the medians in the other populations ranged from 76 to 124 days. The two cohorts with the fewest number of unique codes were the type 2 diabetes controls at NU and VU, where the median number of unique ICD-9 categories, sections and chapters observed were 11, 9, and 6, and 7, 6, and 4, respectively.

Fig. 2 displays the raw prevalence of co-morbidities in several phenotype-site cohort pairs for ICD-9 categories, sections, and chapter using Bland–Altman plots [27], with codes used to define cohorts (250.* codes for type II diabetics; 366.*, 374.*, 385.*, 743.3*, 744.3, 742.3, and 753.0 for senile cataracts; 440, 440.2, 433.*, 433.*, 434.*, 435.*, 436.*, 437.*, 438.*, 441.*, 442.*, 443.*, and 444.* for peripheral arterial disease) having been removed. While these plots

Table 1
Demographic characteristics of the nine eMERGE cohorts under study between January 1, 2001 to December 31, 2007.

	GHC-CAT	MAR-CAT	MAY-CON	MAY-PAD	NU-CON	NU-T2D	VU-CON	VU-QRS	VU-T2D
N	2217	2614	1181	972	850	672	2236	1055	5273
Ethnicity									
African American	0.04	0	0	0	0.08	0.23	0.09	0.13	0.18
Asian	0.03	0	0	0	0	0	0.01	0.01	0.01
Other[a]	0.01	0	0	0	0.07	0.13	0.02	0.01	0.02
Unknown[b]	0.02	0	0.03	0.02	0	0	0.12	0.01	0.02
White	0.90	0.99	0.96	0.98	0.85	0.64	0.76	0.84	0.77
Female	0.62	0.58	0.43	0.36	0.65	0.52	0.64	0.7	0.53
Born before 1928	0.76	0.38	0.04	0.22	0.01	0.04	0.05	0.03	0.08
Age if born in or after 1928	70 (65, 73)	65 (53, 72)	60 (52, 69)	64 (50, 72)	41 (27, 59)	55 (40, 68)	46 (25, 64)	48 (28, 64)	53 (33, 68)
Years of observation	6.7 (4.6, 6.9)	6.7 (5.8, 6.9)	6.3 (4.5, 6.9)	6.3 (4, 6.9)	5.6 (3.6, 6.8)	6.3 (3.8, 6.9)	5.5 (3.4, 6.7)	6 (3.7, 6.8)	6.3 (3.8, 6.9)
Unique visit days	97 (45, 198)	101 (44, 198)	44 (13, 152)	44 (14, 143)	76 (25, 186)	86 (32, 219)	95 (32, 226)	113 (35, 238)	124 (43, 244)
Total ICD9s	221 (93, 483)	215 (90, 452)	159 (50, 496)	160 (51, 503)	196 (47, 579)	243 (62, 640)	193 (57, 520)	221 (67, 574)	249 (84, 604)
Unique ICD-9s	62 (37, 116)	62 (34, 114)	46 (21, 99)	48 (21, 117)	45 (15, 101)	52 (16, 101)	53 (19, 109)	60 (22, 118)	63 (26, 117)
Unique categories	36 (19, 63)	34 (18, 57)	21 (10, 38)	27 (11, 51)	11 (5, 25)	26 (10, 52)	7 (2, 16)	15 (5, 34)	21 (8, 47)
Unique sections	23 (13, 35)	20 (11, 31)	15 (8, 23)	17 (8, 29)	9 (4, 17)	18 (7, 30)	6 (2, 12)	11 (4, 22)	15 (6, 28)
Unique chapters	11 (7, 13)	10 (7, 13)	9 (6, 11)	9 (5, 12)	6 (3, 10)	9 (5, 12)	4 (2, 8)	7 (3, 11)	8 (4, 12)

[a] "Other" ethnicities include Hispanics, Pacific Islander, American Indians, and individuals reporting multiple ethnicities.

[b] "Unknown" ethnicity indicates that no value for this field was recorded in the EMR. Categorical variables are summarized with proportions and continuous variables are summarized with, 50th (10th, 90th) percentiles.

have limitations since they are not adjusted for demographic and other characteristics, they demonstrate interesting patterns. The common site – case versus control plots (NU-T2D versus NU-CON and MAY-PAD versus MAY-CON) in the first two rows of panels show that the cases tend to exhibit a higher prevalence of co-morbidities than their associated controls, though this is more pronounced in the NU plots than in the MAY plots. Chapter level rates between MAY-PAD and MAY-CON appear reasonably similar to one another while even at this highly aggregated level of summarization the NU-T2D cohort tends to exhibit higher rates of morbidities than does its control cohort. The lower two rows of plots display common phenotypes compared across different sites (NU-T2D versus VU-T2D and MAR-CAT versus GHC-CAT). The co-morbidity profiles in these pairs of cohorts are more similar to one another than in the upper two panels. NU-T2D patients tend to experience slightly higher rates of morbidities than VU-T2D patients, though MAR-CAT and GHC-CAT populations appear comparable to one another except in one morbidity category (indicated by the outlying, uppermost point in each of the plots in the bottom row).

Figs. 3 and 4 display the results from stage 1 of the analysis protocol. They show the adjusted odds ratios based on multiple logistic regression models for the 66 ICD-9 sections, ordered alphabetically by ICD-9 chapter and then by section. The size of the plotting points is inversely related to the confidence interval length, although we limited the size of points when confidence intervals were tight, and "X" denotes a very large odds ratio with the lower confidence bound being greater than 20. In Fig. 3, we show within-site, case versus control comparisons at NU, VU, and MAY, and in Fig. 4 we show two, same-phenotype, different-site comparisons (GHC-CAT versus MAR-CAT and NU-T2D versus VU-T2D), and a different-phenotype, different-site comparison (MAY-PAD versus VU-QRS). Consistent with Fig. 2, morbidity profiles in the cohorts with the same phenotype, but at different sites (Fig. 4a and b), are more similar to one another than cases versus controls at the same site (Fig. 3a–c) and different phenotypes at different sites (Fig. 4c) as odds ratios tend to be closer to one.

Figs. 3 and 4 highlight important patterns of differences between pairs of cohorts. For example, while the GHC-CAT and MAR-CAT populations appear to have similar profiles (Fig. 4), we observe that ICD-9 sections "neoplasms of uncertain behavior" and "dislocation" occur at higher rates at GHC than at MAR, and section "other metabolic and immunity disorders" occurs at a much higher rate at MAR than at GHC. This was also observed in Fig. 2. Compared with their controls, adjusted co-morbidity risk was higher for NU-T2D and VU-T2D cohorts over the range of ICD-9 sections, though this result appear less pronounced for the MAY-PAD versus MAY-CON comparison. Fig. 4c shows that the MAY-PAD cohort tended to exhibit higher rates of nervous system and (as expected) circulatory system disorders than the VU-QRS cohort though the opposite was true for neoplasms.

Simple numerical summaries describing differences among populations are complementary and sometimes preferred to the graphical depictions of individual differences, such as in Figs. 3 and 4. The MDI from stage 2 of the analysis protocol for select cohort pairs are shown in Table 2 for ICD-9 sections within chapters and then over the range of all sections. Among the pairwise comparisons, the two same-phenotype, different-site cohorts appeared most similar to one another with an overall MDI of 0.47 for the CAT cohorts (column 5) and 0.44 for the T2D cohorts (column 6). While these values imply non-trivial differences between the cohorts with the same phenotypes at different sites, it is worth noting that the overall MDI for NOR-T2D versus VAN-T2D is just over half the size of the MDI for VAN-T2D versus VAN-CON (MDI = 0.80) and for NOR-T2D versus NOR-CON (MDI = 0.82). That is, the impact of type II diabetes on the overall morbidity profile is approximately 80% larger than the impact of site. Focusing further on ICD-9 sections in the "Endocrine, metabolic and nutritional immunity" ICD-9 chapter, the impact of type II diabetes is at least 150% larger than the impact of site, where MDI is equal to 2.10, 2.18, and 0.83 for VAN-T2D versus VAN-CON, NOR-T2D versus NOR-CON, and NOR-T2D versus VAN-T2D, respectively. ICD-9 sections within the "musculoskeletal system and connective tissue" ICD-9 chapter appeared to be least associated with sites and phenotypes as MDI values ranged from 0.16 for MAY-PAD versus MAY-CON, to 0.50 for the VU-T2D versus VU-CON.

It should be noted that with finite samples, the MDI measure would be non-zero even when cohorts are randomly sampled from the same populations. However, with large samples such as those discussed here, under random sampling from a single population,

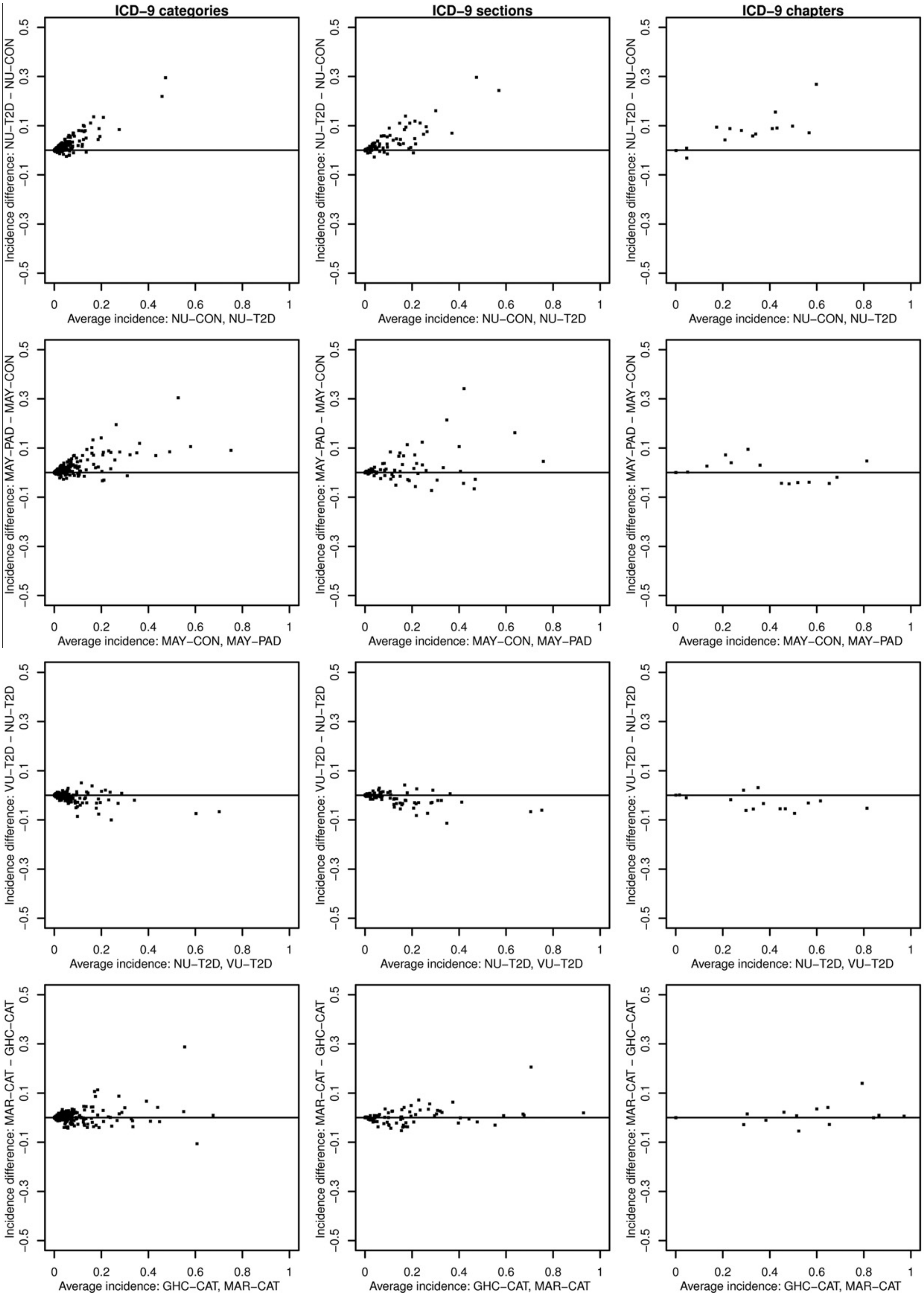


Fig. 2. Bland–Altman plots comparing unadjusted rates of ICD-9 categories, sections, and chapters for pairs of populations.

it will be very close to zero. We conducted all analyses having repeatedly and randomly reassigned cohort identifiers (e.g., using a Monte-Carlo based randomization approach to simulate random samples from a single population). After rounding to the nearest hundredth, none of the values corresponding to those shown in Table 2 exceeded 0.02.

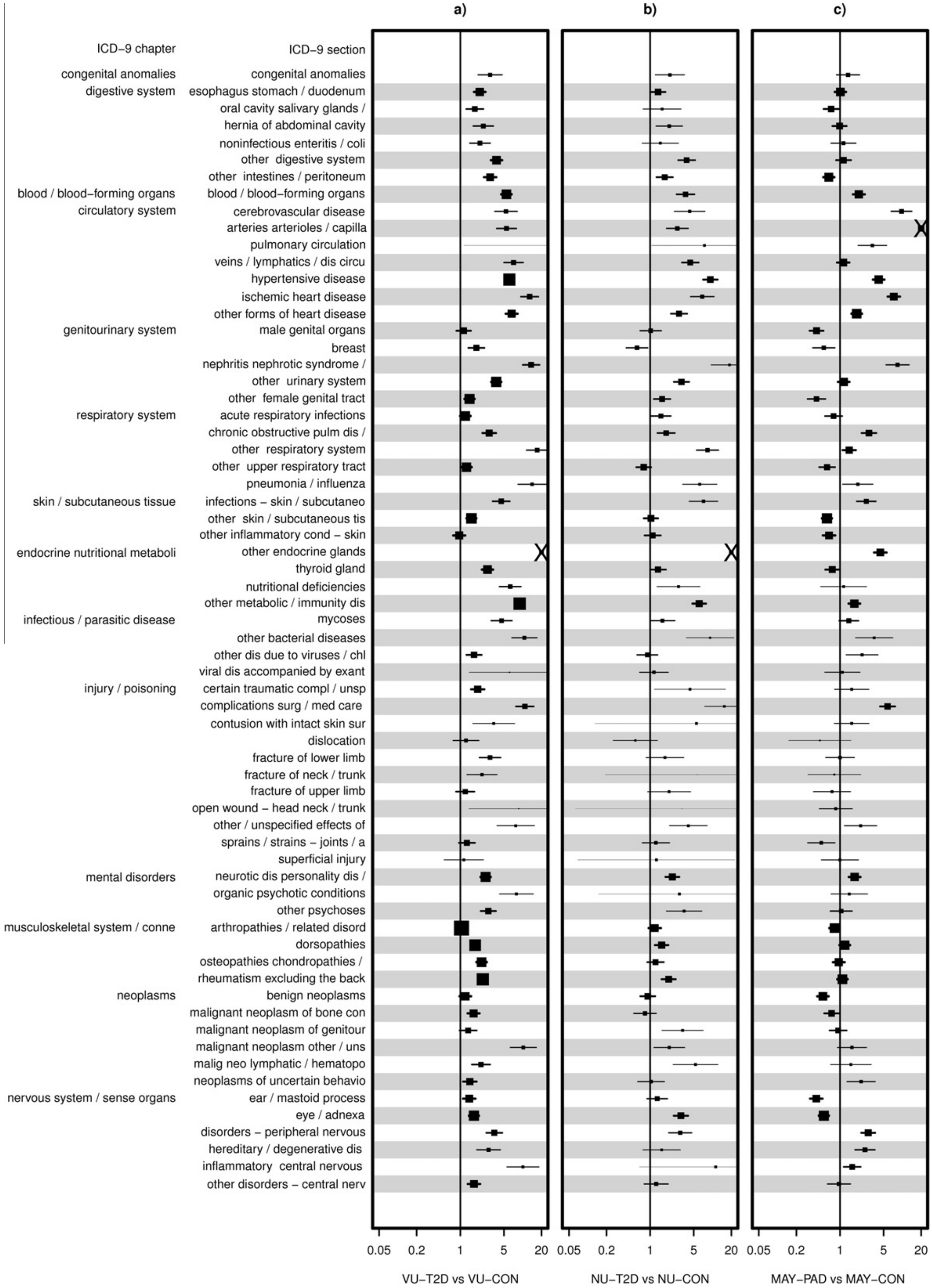


Fig. 3. Adjusted odds ratios, based on stage 1 logistic regression analyses, comparing VU-T2D to VU-CON, NU-T2D to NU-CON, and MAY-PAD to MAY-CON. The symbol “X” denotes an extremely high odds ratio whose lower confidence limit exceeds 20.

5. Discussion

We have proposed a general two-stage analysis approach for systematic characterization of co-morbidity profile differences between cohorts derived from EMRs. The strategy involves regression modeling over a range of ICD-9 sections, graphical displays of results, and summarization of the differences with the MDI for broader insights. Results from first and second stage analyses are

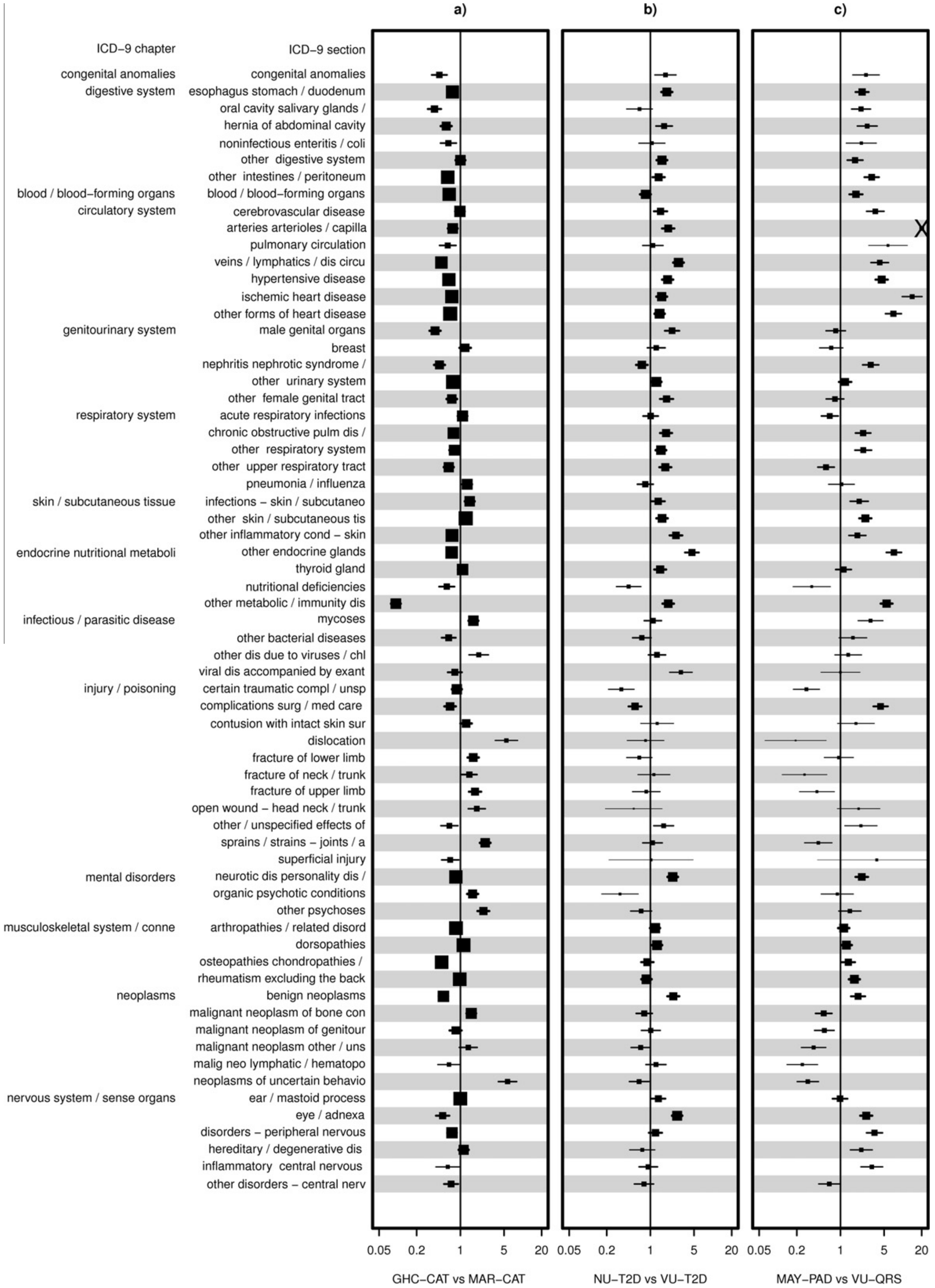


Fig. 4. Adjusted odds ratios, based on stage 1 logistic regression analyses, comparing GHC-CAT to MAR-CAT, NU-T2D to VU-T2D, and MAY-PAD to VU-QRS. The symbol "X" denotes an extremely high odds ratio whose lower confidence limit exceeds 20.

complementary, and the breadth of the co-morbidities one chooses to examine depends on study objectives. If the objective is to characterize dissimilarity broadly (e.g., comparing the differences between two hospitals or finding the "nearest neighbor" between two cohorts) then a diverse range of morbidities should be considered. However, if the objective is to anticipate analytical challenges to a multicenter study (e.g., variance inflation or power reduction due to among site heterogeneity) where the target phenotype has been identified but has not yet been ascertained, then the range of morbidities to consider should be narrower and should be related to the target phenotype. The MDI is on the same scale as parameters in logistic regression analyses, and so it has an intu-

Table 2
Morbidity Dissimilarity Index for cohort pairs.

Chapter	VU-T2D vs VU-CON	NU-T2D vs NU-CON	MAY-PAD vs MAY-CON	GHC-CAT vs MAR-CAT	NU-T2D vs VAN-T2D	VAN-QRS vs MAY-PAD
Congenital anomalies	1.10	0.73	0.30	0.77	0.55	0.94
Digestive system	0.88	0.70	0.24	0.39	0.39	0.74
Diseases blood and blood-forming organs	1.70	1.30	0.70	0.41	0.18	0.57
Diseases of the circulatory system	1.71	1.35	1.44	0.34	0.49	1.70
Diseases of the genitourinary system	1.02	0.94	0.88	0.48	0.43	0.51
Diseases of the respiratory system	0.95	0.92	0.61	0.27	0.38	0.64
Diseases of the skin and subcutaneous tissue	0.62	0.68	0.56	0.29	0.52	0.69
Endocrine nutritional metabolic immunity	2.10	2.18	0.91	1.20	0.83	1.47
Infectious and parasitic diseases	1.32	0.65	0.59	0.48	0.51	0.64
Injury and poisoning	0.95	1.20	1.07	0.56	0.53	1.15
Mental disorders	0.84	0.81	0.45	0.47	0.82	0.63
Musculoskeletal system and connective tissue	0.50	0.34	0.16	0.36	0.20	0.26
Neoplasms	0.58	0.51	0.51	0.66	0.56	0.81
Nervous system and sense organs	0.73	0.81	0.72	0.28	0.61	0.82
Across all ICD-9 sections	0.80	0.82	0.66	0.47	0.44	0.75

itively appealing interpretation. It can also be exponentiated if one wishes characterize dissimilarity with odds ratios.

In the eMERGE study analysis, we found that cohorts with the same phenotypes at different institutions appeared to have more similar morbidity profiles than those representing different phenotypes, providing some reassurance for the planned network projects. We intend to perform this analysis on many eMERGE projects prior to their implementation, as results and implications will depend upon the phenotype. As more of the phenotype defined populations become available, these and other data will better inform the development of general guidelines for how ‘similar’ populations should be for pooled genetic or clinical analysis. While the MDI can be interpreted as a measure of dissimilarity on the scale of log odds ratios, its aptness for proactively capturing potentially heterogeneous site-specific genotype-phenotype associations depends on a number of data features and perhaps most importantly on the strength of the relationship between the morbidities that comprise it and the phenotype of interest. The stronger the relationship, the more likely it is to be useful. That being said, it can only be used as a guide since morbidity profile dissimilarity does not capture genotype-phenotype association heterogeneity. As an area for future research, we will explore various data features that impact the utility of the MDI for this aim. We will also explore the utility of formally incorporating domain structure (i.e., the ICD-9 taxonomy) into the calculation of the overall MDI. In our two-stage approach, we acknowledged domain structure explicitly by organizing Figs. 3 and 4 by ICD-9 chapters and by calculating chapter-specific MDIs; however the domain structure was not implicit in the calculation of the overall MDI. The formal incorporation of this structure will effectively involve a reweighting of ICD-9.

This analytical protocol is not limited to the ICD-9 coding system and could be used for other classification schemes, such as CPT codes, medications given, or NLP-derived disease codes mapped to controlled terminologies such as SNOMED or the UMLS. Using NLP may improve recall and precision of disease identification [28,29]. One challenge, if mapping to a vocabulary such as the UMLS, would be to aggregate codes at an appropriate level. For instance, as discussed earlier, we found that performing the tests of associations using ICD-9 category codes (904 unique codes) provided insufficient counts of patients with each code to allow for statistical analysis. Thus, a large percentage of possibly important codes would have been removed from the analysis.

There are several limitations of this study. There are a number of known problems with ICD-9 codes for diagnosis, including false positives and false negatives [30]. At eMERGE network institutions, professional coders typically entered inpatient codes, while outpatient codes resulted from direct physician entry. Invalid or incorrect codes are often entered, either from memory or from prepopulated lists (e.g., a type 1 diabetes code when a type 2 code is intended). Codes that are difficult to find or that do not lead to significant reimbursement may be excluded. Some institutions arbitrarily limit the number of codes stored in their data warehouse from a particular visit, while others do not, and some data warehouses include both incorrect and corrected codes. The ICD-9 hierarchy itself is not optimal for phenotypic analysis, since it is designed and maintained to support administrative and billing operations. In addition, coding practices can vary among practitioners within institutions and between institutions. We considered only diagnosis codes and demographics in our comparisons, and due to age truncation, there is likely to be residual confounding. Other health information, such as medication information and procedures received, are important markers of the veracity and severity of disease and if available could also be included in analyses. Finally, we did not utilize disease onset times. It would be very interesting to conduct analyses that consider morbidity timing and morbidity coding in relation to disease onset times. For example, one could examine how coding practices change from before disease onset to after disease onset, or one could examine coding trends leading up to the time of disease onset.

Future clinical and genomic research will benefit from deriving samples from diverse data repositories. The ability to investigate rare diseases for genomic and environmental influences will require aggregation of samples from multiple repositories. We present an initial attempt to highlight and quantify the non-random influences of geographic and provider practices to inform analysis of such data. More research is needed to study the certainty of ICD-9 codes and use of other resources to improve the accuracy of comorbidity assessment and severity.

References

[1] Herzig SJ, Howell MD, Ngo LH, Marcantonio ER. Acid-suppressive medication use and the risk for hospital-acquired pneumonia. JAMA 2009;301(20): 2120–8.

[2] Klompas M, Haney G, Church D, Lazarus R, Hou X, Platt R. Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. PLoS ONE 2008;3(7):e2626.

[3] Dean BB, Lam J, Natoli JL, Butler Q, Aguilar D, Nordyke RJ. Review: use of electronic medical records for health outcomes research: a literature review. Med Care Res Rev 2009 Dec;66(6):611–38.

[4] Denny JC, Arndt FV, Dupont WD, Neilson EG. Increased hospital mortality in patients with bedside hippus. Am J Med 2008;121(3):239–45.

[5] Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007;13(6 Part 1):281–8.

[6] Seyfried L, Hanauer DA, Nease D, Albeiruti R, Kavanagh J, Kales HC. Enhanced identification of eligibility for depression research using an electronic medical record search engine. Int J Med Inform 2009 Dec;78(12):e13–18.

[7] Burton PR, Hansell AL, Fortier I, Manolio TA, Khoury MJ, Little J, et al. Size matters: just how big is BIG?: quantifying realistic sample size requirements for human genome epidemiology. Int J Epidemiol 2009;38(1):263–73.

[8] Manolio TA. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI's office of population genomics. Pharmacogenomics 2009 Feb;10(2):235–41.

[9] DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986 Sep;7(3):177–88.

[10] Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40(5):373–83.

[11] Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care 1998 Jan;36(1):8–27.

[12] van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. Med Care 2009 Jun;47(6):626–33.

[13] Knaus WA, Zimmerman JE, Wagner DP, Draper EA, Lawrence DE. APACHE-acute physiology and chronic health evaluation: a physiologically based classification system. Crit Care Med 1981;9(8):591–7.

[14] Groll DL, To T, Bombardier C, Wright JG. The development of a comorbidity index with physical function as the outcome. J Clin Epidemiol 2005 Jun;58(6):595–602.

[15] Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. J Biomed Inform 2008 Dec;41(6):882–8.

[16] Cavalli-Sforza LL, Edwards AW. Phylogenetic analysis models and estimation procedures. Am J Hum Genet 1967;19(3 Pt 1):233–57.

[17] Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet 2003 Mar; 33(Suppl):266–75.

[18] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38(8):904–9.

[19] Lee C, Abdool A, Huang C. PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 2009;10(Suppl 1):S73.

[20] The eMERGE network. [cited 2009 9/13]; Available from: http://www.gwas.net.

[21] Ioannidis JPA, Trikalinos TA, Khoury MJ. Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases. Am J Epidemiol 2006;164(7):609–14.

[22] UMLS Knowledge Source Server. [cited 2007 July 3]; Available from http://umlsks/nlm/nih.gov/kss/.

[23] Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. Bioinformatics 2010;26(9):1205–10.

[24] Ritchie MD, Denny JC, Crawford DC, Ramirez AH, Weiner JB, Pulley JM, et al. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 2010;86(4):560–72.

[25] Harrell F. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001.

[26] Efron B. An introduction to the bootstrap. New York: Chapman and Hall; 1993.

[27] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1(8476):307–10.

[28] Elkin PL, Ruggieri AP, Brown SH, Buntrock J, Bauer BA, Wahner-Roedler D, et al. A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM. Proc AMIA Symp 2001:159–63.

[29] Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc 2008:404–8.

[30] Aronsky D, Haug PJ, Lagor C, Dean NC. Accuracy of administrative data for identifying patients with pneumonia. Am J Med Qual 2005 Dec;20(6):319–28.