The Annals of Applied Statistics
2017, Vol. 11, No. 3, 1349–1374
DOI: 10.1214/17-AOAS1028

COMPARING HEALTHCARE UTILIZATION PATTERNS VIA GLOBAL DIFFERENCES IN THE ENDORSEMENT OF CURRENT PROCEDURAL TERMINOLOGY CODES[1]

BY XU SHI*, HRISTINA PASHOVA† AND PATRICK J. HEAGERTY*

*University of Washington** *and Axio Research*†

The linkage of electronic medical records (EMR) across clinics, hospitals, and healthcare systems is opening new opportunities to evaluate factors associated with both individual treatment benefit and potential harm. For example, the FDA Sentinel initiative seeks to create a surveillance network with over 100 million patient lives (Behrman et al. [*N. Engl. J. Med.* **364** (2011) 498–499]), while PCORnet has created multiple networks that include linked electronic medical records from geographic regions such as entire cities or states, with the ultimate goal of facilitating comparative effectiveness research (Collins et al. [*Journal of the American Medical Informatics Association* **4** (2014) 576–577]). However, one key challenge to the use of electronically assembled cohorts is the potential for variation in both the choice of specific healthcare procedures and coding practices due to differences in patient populations and/or financial incentives within care delivery networks. In order to explore variation in patient care or procedure coding, we review and develop statistical methods that can permit testing and estimation of subgroup differences in code assignments. We focus on Current Procedural Terminology (CPT) codes which are used in a standardized fashion to capture patient treatment details and to record medical histories, but the methods we develop can be used for any structured EMR data. We specifically study testing procedures that can be valid for comparing both rare and common counts as routinely encountered with medical procedure codes, and we transfer methods from studies of genetic association. Hierarchical structure in terms of both thematically grouped medical codes and provider-level clustering adds unique complexity to the analysis of EMR data. We detail penalized regression methods unifying estimation and inference to leverage the hierarchical structure and stabilize rate ratio estimates for rare procedures. We also expand inference methods to account for potential within provider correlation of patient utilization. We illustrate methods comparing the endorsement of CPT codes for subjects enrolled in a back pain cohort study where interest is in the differences across recruitment centers in the use of CPT codes (Jarvik [*BMC Musculoskelet Disord.* **13** (2012)]).

1. Introduction. In the United States, the use of electronic medical records (EMR) is now incentivized due to the 2009 enactment of the Health Information Technology for Economic and Clinical Health (HITECH) Act. Large scale EMR

Received November 2016; revised January 2017.
[1]Supported by Grant NIH UL1 TR000423, NIH R01 HL072966, and AHRQ R01 HS22972.
Key words and phrases. Electronic medical records, hierarchical structure, dynamic graphics.

data will open new opportunities for research to improve patient care and the health of the public. Current national research efforts include linking EMR to conduct pharmacosurveillance (e.g., FDA Sentinel) and assembling large clinical populations for comparative effectiveness research (e.g., PCORnet). One key element of EMR data is the recording of patient treatment history via the Current Procedural Terminology (CPT) coding system. CPT codes are five-character codes describing medical services and procedures, and are used for patient management and billing. CPT codes provide a standardized description that allows communication across providers and systems, and facilitate identification of clinical information for comparative effectiveness research.

A natural research question is whether different subgroups of patients have different healthcare utilization patterns, and interest may lie in the entire spectrum of all potential services. A motivating example is the Back pain Outcomes using Longitudinal Data (BOLD) project, which enrolled 5239 patients who are 65 years of age or older with back pain from multiple healthcare systems with a primary interest in early imaging [Jarvik (2012)]. In this study, we combine electronic health records across sites, and there is need to assess potential data quality concerns by comparing codes between healthcare systems. Ultimately, the BOLD study compared healthcare utilization for propensity score matched patients with and without early radiologic imaging. We were interested in effect of early imaging on down stream healthcare utilization in terms of a summary measure of total spine-related procedures, but we also wanted to examine the full set of individual CPT codes and overall utilization. There is an emerging need to develop inference methods that are tailored for the electronic medical records context. Recent federal healthcare and research initiatives are incentivizing the routine collection of population scale data, and statistical methods are needed for evaluation of data quality and for high-throughput comparison of utilization for select patient subsets.

However, the use of EMR data for research comes with several challenges. First, a unique aspect of EMR data is the organizational structure where individual patients are typically nested within providers who may have unexplained variation in their treatment patterns that induce correlation in utilization indicators for their patient panel. Second, an important characteristic of CPT codes is the hierarchical structure in coding taxonomy: multiple CPT codes may represent similar or related procedures. According to the Clinical Classifications Software (CCS) for Services and Procedures, CPT codes are naturally collapsible into 244 clinically meaningful groups, which define major categories of procedures. The CCS-Services and Procedures taxonomy is a part of the Healthcare Cost and Utilization Project (HCUP), a Federal-State-Industry partnership sponsored by the Agency for Healthcare Research and Quality. See https://www.hcup-us.ahrq.gov/toolssoftware/ccs_svcsproc/ccssvcproc.jsp for details on the CCS classification. Inference can be made at either the code or the group of codes level, and shrinkage methods may be desired to borrow strength from similar codes. Third, in practice, both investigation of data quality and evaluation of overall utilization imply the

need to perform inference for thousands of codes, and robust methods that can be applied to both common codes and rare procedures are needed. In this paper, we consider testing and estimation methods for quantifying the significance and magnitude of differences in the delivery of all possible procedures and services between two cohorts of patients. Our proposed methods are tailored to the unique and increasingly important context of healthcare delivery system generated observational data with the above mentioned challenges: patient clustering, CPT code grouping, and coexistence of common and rare procedures.

The paper is organized as follows. Section 2 details testing for differences in healthcare utilization. Section 2.1 focuses on code-wise testing procedures for both common and rare codes. Section 2.2 considers testing of procedures defined by a group of codes. Section 2.3 discusses methods that account for provider-level clustering. In Section 3, we propose rate ratio estimation and inference methods. Section 3.1 details a ridge regression model which takes advantage of the hierarchical structure of CPT codes and stabilizes estimates of rare procedures. In Section 3.2, we provide inference method accounting for shrinkage bias. Section 3.4 considers inference with provider-level clustering. Both methods in Sections 2 and 3 allow for confounding adjustment. In Section 4, we conduct simulation studies to evaluate the performance of the code-wise and group-wise tests, as well as the inference for ridge regression. We also study the influence of provider-level clustering on testing procedures and the performance of methods that accounts for within-provider correlation. In Section 5, we illustrate the methods by analyzing EMR data from the BOLD study. We evaluate differences in CPT codes assigned among patients from two healthcare systems: Kaiser Permanente in Northern California and Henry Ford Health System in Detroit. For healthcare systems or clinics within systems, the benchmarking of one site against a reference site is an important part of revealing variation that may require attention in order to align delivery decisions with clinical guidelines or to potentially reduce cost. Therefore, differences in healthcare utilization across the entire set of codes are generally important to evaluate for both delivery assessment and research purposes. We develop and illustrate graphical tools that compare patient subgroups across the full spectrum of procedures and services for exploratory research using large scale healthcare data. We close with a discussion in Section 6.

2. Testing for utilization of procedures defined by a CPT code or a block of codes. For simplicity, we consider patients' visits over a fixed time period such as one year, although methods that characterize rates of code endorsement can easily handle variable follow-up time at the patient level by weighting the outcome with the inverse of patient-specific length of follow-up. Let n_s, $s = 0, 1$, denote the total subjects in cohort s, and without loss of generality, we assume one year of follow-up for each subject. For a specific procedure described by CPT code c, we take two approaches to comparing utilization rates across cohorts. First, we consider an outcome based on a count of how many times the procedure was delivered to

TABLE 1

Summary of two-sample testing options for CPT count and binary Data. The notation Λ denotes the likelihood ratio test statistic

Distribution	**Likelihood ratio test**	**Conditional exact test**
Poisson	$Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$Y_1^c \mid Y_0^c + Y_1^c$ $\overset{H_0}{\sim} \text{Binomial}(n = Y_0^c + Y_1^c, p = \frac{n_1}{n_0+n_1})$
Negative Binomial	$Y_{si}^c \sim \text{Neg-Bin}(p = \frac{\lambda_s^c}{\lambda_s^c + \frac{1}{\phi^c}}, r = \frac{1}{\phi^c})$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$\Pr(Y_0^c = y_0^c, Y_1^c = y_1^c \mid Y_0^c + Y_1^c)$ $\overset{H_0}{=} \frac{\binom{y_0^c + \frac{n_0}{\phi^c} - 1}{y_0^c} \cdot \binom{y_1^c + \frac{n_1}{\phi^c} - 1}{y_1^c}}{\binom{y_0^c + y_1^c + \frac{n_0+n_0}{\phi^c} - 1}{y_0^c + y_1^c}}$
Binomial	$Z_{si}^c \sim \text{Bernoulli}(p_s^c)$ $-2\log(\Lambda) \overset{H_0}{\to} \chi^2$	$Z_1^c \mid Z_0^c + Z_1^c$ $\overset{H_0}{\sim} \text{Hypergeometric}(N = n_0 + n_1,$ $n = Z_0^c + Z_1^c, K = Z_1^c)$
Semiparametric	t-test	

patient i in cohort s over the year, denoted as Y_{si}^c, where $s = 0, 1$, $i = 1, \ldots, n_s$, $c = 1, \ldots, C$. Second, for certain scientific questions we may only be interested in whether the procedure was ever endorsed for a subject, and we may choose to dichotomize the count data into any/none outcomes $Z_{si}^c = \mathbb{1}(Y_{si}^c > 0)$, as an indicator of whether patient i was assigned code c in any visits over the year.

2.1. *Testing for code-specific patient utilization.* Our interest in CPT code-wise inference requires selection of a testing strategy that can be valid for both common and rare codes, and under potential overdispersion. However, in practice, there is little guidance on how to choose appropriate tests. In this section, we discuss various two-sample testing strategies that are candidates for the evaluation of variation in code endorsement rates across cohorts for count and binary data with a goal of characterizing the applied options, summarized in Table 1. In Section 4, we perform numerical studies to illustrate the performance of testing options for a range of rate parameters that may be expected from CPT data. Although we focus on a crude comparison, adjustment for covariates can be achieved through stratification or matching on the propensity score, which is the probability that a patient belongs to a cohort given the observed confounders [Rosenbaum and Rubin (1983, 1984)].

2.1.1. *Count outcome.* For count data, a natural model is the Poisson distribution characterized by a rate parameter, λ_s^c, with $Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$, $i = 1, \ldots, n_s$. In cohorts where there are both relatively healthy and extremely ill patients, there will be overdispersion, and the simple Poisson mean–variance relationship will not

hold. In this situation, use of negative binomial distribution provides one model-based generalization of the Poisson assumption. The negative binomial model contains a rate parameter and an additional parameter that characterizes overdispersion: $Y_{si}^c \sim$ negative binomial$(p = p_s^c = \frac{\lambda_s^c}{\lambda_s^c + \frac{1}{\phi^c}}, r = 1/\phi^c)$, $i = 1, \ldots, n_s$, where ϕ^c is the code-specific overdispersion parameter which is shared by patients across cohorts. The mean and variance are parameterized as $\mathrm{E}[Y_{si}^c] = \lambda_s^c$ and $\mathrm{Var}[Y_{si}^c] = \lambda_s^c(1 + \lambda_s^c \phi^c)$. We wish to test whether code endorsement varies by cohort, that is, H_0: $\lambda_0^c = \lambda_1^c$.

In the EMR setting, the number of patients is usually quite large (thousands or greater) and large sample approximations should be valid. Therefore, if the model assumption was valid, then a likelihood ratio test (LRT) using observations from each patient could provide inference regarding coding rates across cohorts. However, rare codes can lead to low expected cell counts, which can lead to a poor χ^2 approximation for the null distribution of the LRT. In this case, as an exact alternative, we can collapse patient-level information within a cohort by computing the total counts $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c$ and apply a conditional exact test (ET) which calculates a conditional probability that does not depend on large sample approximations [Przyborowski and Wilenski (1940), Robinson, McCarthy and Smyth (2010)].

While use of the negative binomial model allows a partial decoupling of the mean and the variance, it may not provide valid inference when the true data generating mechanism is not adequately characterized by a simple overdispersed count model. Alternatively, with large sample sizes and any underlying distribution, we can use the two-sample t-test as a semiparametric method for testing.

2.1.2. *Any/none outcome*. Use of endorsement rates will often be the appropriate strategy for answering scientific questions about variation in utilization. However, for certain codes such as recommended annual screening measures or vaccinations, it may be desirable to simply analyze the count data as a derived binary outcome since the clinical significance is indicated by any endorsement of the code. An any/none outcome indicates whether a patient was ever assigned a code during his or her visits over the year. It can be modeled as $Z_{si}^c \sim$ Bernoulli(p_s^c) for a patient, and $Z_s^c = \sum_{i=1}^{n_s} Z_{si}^c \sim$ Binomial(n_s, p_s^c) for a cohort. Our goal is to test whether the probability of assigning a CPT code varies by cohort, that is, H_0: $p_0^c = p_1^c$. When the requirements of the χ^2 approximation are met for the LRT, we use the Binomial LRT. When the expected cell counts are too small, we can use the conditional ET assuming Binomial model, which is the well-known Fisher's ET for a two-by-two table constructed using cohort level data.

We summarize standard asymptotic LRTs and ETs for count and binary data in Table 1, and provide detailed reviews in Supplement A of the supplementary materials [Shi, Pashova and Heagerty (2017)]. In summary, key practical issues include: whether to adopt asymptotic or exact tests; whether to consider a count or indicator outcome; and whether or how to account for overdispersion in CPT counts.

2.2. *Testing for block-specific patient utilization*. It has been shown that for a specific procedure, the level of agreement among coders and agencies in assigning CPT codes can be poor [Bentley et al. (2002), Holt, Warsy and Wright (2010), King, Lipsky and Sharp (2002), King, Sharp and Lipsky (2001)]. That is, physicians might use different codes to describe the same procedure since multiple codes can be appropriate for a certain general procedure. For example, bilateral screening mammography, according to the imaging technology, can be coded using CPT code 77057 which is labelled "Screening mammography, bilateral (2-view film study of each breast)", or can be recorded using CPT code G0202 labelled as "Screening mammography, producing direct digital image, bilateral, all views". Therefore, code-level analysis may detect variation that is not reflective of meaningful practice variation, and analysis at a "code group" level may be more appropriate. According to the CCS, CPT codes can be collapsed into groups. We call such a group of codes a "block". Procedures can be compared at CCS block-level where finer scale differences in procedure coding may not indicate an overall difference, so that the comparison is not sensitive to physicians' choice of codes within a block.

Suppose that there are C total codes that can be categorized into B blocks. Let $S(b)$ be the set of codes that belong to block b, and let C_b denote the number of codes in $S(b)$, such that $\sum_{b=1}^{B} C_b = C$. Given the hierarchical structure of medical codes, we use $Y_{si}^{bc} = Y_{si}^{c}$ to emphasize that each code c belongs to a certain block b. Therefore, the count vector Y_{si}^{bc} for $b = 1, \ldots, B$ and $c = 1, \ldots, C_b$ corresponds to one observation or row of data associated with patient i in cohort s. In the following sections we detail testing methods that can be used to make inference at the block level.

2.2.1. *Burden test*. A simple testing procedure parallels methods used for genomic data that have been termed "burden tests", since the total number of endorsements within a block (i.e., total burden) is used as the basis for testing [Madsen and Browning (2009), Morgenthaler and Thilly (2007), Morris and Zeggini (2010)]. Using this approach we apply the code-wise testing methods in Section 2.1 to block-level summaries. Specifically, for a procedure defined by block b, $Y_{si}^{b} = \sum_{c \in S(b)} Y_{si}^{bc}$ summarizes the assignment of the block-specific procedure to patient i in cohort s over the year, and $Z_{si}^{b} = \mathbb{1}(Y_{si}^{b} > 0)$ describes whether the procedure was assigned to patient i at any visits over the year. Within genomic research, the burden test is a group-wise association test that potentially increases power by combining genetic counts within regions or genes [Wu et al. (2011), Lee et al. (2012)]. In our context, when codes within a given block are consistently used more frequently in one cohort, the burden test accumulates individual code effects to increase power. Such a strategy is especially important when dealing with rare codes. However, for codes that have inconsistent distributions, combining their effects when they may be in opposite directions across the comparison groups can

lead to cancellation of potentially meaningful variation and null test results. Thus, the burden test might diminish code-level effects when they are aggregated. On one hand, such aggregation ensures that the burden test is insensitive to code substitution, but on the other hand, it may decrease power to detect a meaningful code-level variation.

2.2.2. *Sequence kernel association test*. We use the sequence kernel association test (SKAT) as a complement to the burden test [Wu et al. (2011)]. This test is based on a mixed model framework that was developed to collectively test for the association between a set of genetic variants and a phenotype. We treat CPT codes within a block as analogues to genetic variants within a region, and the cohort of a patient as the dichotomous phenotype to allow testing for groups of procedure codes.

For each block b, we consider the logistic regression model for the phenotype, that is, for cohort s

$$\operatorname{logit}\bigl(\Pr(s=1)\bigr) = \alpha_0 + \sum_{c\in S(b)} \beta_c Y_{si}^{bc},$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{C_b}) \sim N(\mathbf{0}, \sigma^2 \boldsymbol{W})$, where $\boldsymbol{W} = \operatorname{diag}(w_1, \ldots, w_{C_b})$ is a weight matrix. Testing for the null hypothesis $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to the variance component test for $H_0 : \sigma = 0$. Let $\hat{\boldsymbol{\mu}}_0$ denote the expectation of outcome $\Pr(s=1)$ under the null. The score test statistic is $\mathbf{Q} = (\mathbf{s} - \hat{\boldsymbol{\mu}}_0)^T \mathbf{K} (\mathbf{s} - \hat{\boldsymbol{\mu}}_0)$, where $\mathbf{K} = \mathbf{YWY}^T$ is a weighted kernel, and $\mathbf{Y}$ is an $(n_0 + n_1) \times C_b$ matrix whose elements are Y_{si}^{bc}. It is important to note that SKAT can be applied to either a matrix of binary data $\mathbb{1}(Y_{si}^{bc} > 0)$ similar to single nucleotide polymorphism (SNP) data, or count data Y_{si}^{bc} similar to RNA-seq data. In addition, adjustment for confounders can be achieved through fitting the model:

$$\operatorname{logit}\bigl(\Pr(s=1)\bigr) = \alpha_0 + \alpha Z_i + \sum_{c\in S(b)} \beta_c Y_{si}^{bc},$$

where Z_i denotes the set of covariates for patient i [Wu et al. (2011)].

SKAT collectively tests for the association between the endorsement of codes within a block and the cohort. It can increase power by summarizing over multiple codes but does not require that associations are all in the same direction. However, when the code effects are truly in the same direction, simulation studies showed that burden tests may have higher power [Wu et al. (2011)]. Thus, we recommend that both the burden test and the SKAT are used with awareness of the types of departures that would likely be detected.

2.3. *Provider-level clustering*. In addition to patient-level variation, there may be provider-level variation in use of CPT codes. Specifically, providers may have individual preferences in their typical choice of treatment paths, which ultimately introduces correlation between patients treated by the same provider.

The impact of ignoring correlation for clusters of patients generally lies in estimation of the standard errors or test statistic variances. For two-sample testing procedures employed in code-wise comparisons and block-wise burden test, we consider using a generalized estimating equation (GEE)-type sandwich variance estimator with working independence covariance matrix in a z-statistic to replace the t-test [Diggle et al. (2002)]. For SKAT, Qi et al. (2015) considered expanding SKAT for longitudinal data and developed the longitudinal kernel machine regression (L-KM) method. In our context, the L-KM essentially adds a provider-level random intercept to introduce correlation between patients within provider. Compared to SKAT, the variance-covariance matrix in the test statistic is estimated from a mixed model which takes into account correlation within provider.

When the primary care provider is taken to define the cluster, patients are nested within providers and our proposed testing procedure with a GEE-type sandwich estimator can correct the type I error to validly account for any provider level clustering. However, when patients are treated by potentially different providers, there could be multilevel clustering that is not necessarily nested. In this case, variance estimation accounting for such multilevel clustering requires extra attention and additional complexity. One strategy would be a likelihood-based random effects approach with implementation using Bayesian computational techniques to provide valid inference. Alternatively, the GEE-type variance estimator with working independence covariance matrix proposed in our testing procedures can be generalized to account for nonnested clustering, which was introduced in Miglioretti and Heagerty (2004) and Miglioretti and Heagerty (2007). The nonnested sandwich variance calculation is surprisingly simple and easy to generalize to more than two nonnested levels of clustering.

In Section 4.1.2, we conduct simulation studies to investigate the sensitivity/robustness to provider-level clustering of the two-sample tests. We also study the performance of the above tests with appropriate correction for within-provider clustering.

3. Rate ratio estimation and inference using ridge regression. In Sections 2.1 and 2.2, we detailed testing methods that assess the statistical significance associated with the observed difference in healthcare utilization across two cohorts. However, testing provides only a partial characterization of utilization differences, and two further questions are of interest: Which cohort is the frequent user, and how large is the magnitude of difference? Recall that CPT codes may be rarely used (e.g., zero counts can be common) and are nested within blocks. We take advantage of such a hierarchical structure and stabilize estimates of rare codes by using a Poisson regression model [equation (1)] with Ridge penalty [Hoerl and Kennard (1970)] that estimates code-specific cohort effects for all codes simultaneously, and exploits potential similarity of endorsement trends within blocks.

3.1. *Rate ratio estimation using hierarchical shrinkage.* Let $Y_{s'i'}^{b'c'}$ denote the number of assignments of code c' within block b' to patient i' in cohort s'. Recall that in Section 2.2 we let $Y_{\cdot}^{bc} = Y_{\cdot}^{c}$ to emphasize that code c belongs to a certain block b. For each patient i' and for each code c' assigned to this patient, define indicators that denote $(s', b', c')_{i'}$ using $\text{cohort}_{1,i'} = \mathbb{1}\{s' = 1\}$, $\text{block}_{b,i'} = \mathbb{1}\{b' = b\}$, and $\text{code}_{c,i'} = \mathbb{1}\{c' = c\}$, where $b = 2, \ldots, B$ and $c \in \{S(b) \setminus c_{\text{ref}}^{b}\}$, with reference levels $s = 0$ for cohorts, $b = 1$ for blocks, and $c = c_{\text{ref}}^{b}$ for codes in block b. Take $t_{i'}$ as the offset to account for potentially different lengths of follow-up across patients, which also defines the rate of code endorsement for patient i' as $\mathrm{E}[Y_{s'i'}^{b'c'}]/t_{i'}$. In addition, we consider confounding adjustment and denote the vector of observed covariates for patient i' as $\boldsymbol{Z}_{i'}$. The model is

$$
\begin{aligned}
&\log\bigl[E\bigl(Y_{s'}^{b'c'}\bigr)\bigr] \\
&\quad = \log(t) + \alpha_0 + \alpha_1 \text{cohort}_1 + \boldsymbol{Z}^{\boldsymbol{T}}\boldsymbol{\theta} \\
&\quad + \sum_{b=2}^{B} \alpha_b \text{block}_b + \sum_{b=2}^{B} \gamma_{b1} \text{block}_b \cdot \text{cohort}_1 \\
&\quad + \sum_{b=2}^{B} \sum_{c \in \{S(b) \setminus c_{\text{ref}}^{b}\}} \eta_c \text{code}_c + \sum_{b=2}^{B} \sum_{c \in \{S(b) \setminus c_{\text{ref}}^{b}\}} \zeta_{c1} \text{code}_c \cdot \text{cohort}_1,
\end{aligned} \tag{1}
$$

where we suppress the notation i' for simplicity.

In this model, the rate ratio comparing cohorts 1 and 0 is determined by three components: the main effect of cohort, α_1; the block-cohort interaction that adds an increment to the overall cohort level, γ_{b1}; and the code-cohort interaction that adds an increment to the overall block level, ζ_{c1}. In other words, the rate ratio on the log scale is defined as

$$\log(\text{RR}) = \alpha_1 + \gamma_{b1} + \zeta_{c1}.$$

The rationale for building a multi-level structure that includes all blocks and codes is to leverage hierarchical shrinkage to control the extent of information sharing and to stabilize rate ratio estimates for rare procedures, with a primary goal of visualization of the effect sizes. A ridge penalty governs the shrinkage of code-specific rate ratio estimates toward the block-level rate ratio, which essentially represents an average over all codes within the same block. In this way, we allow rare codes to borrow information from similar codes within the blocks. We also penalize across blocks to provide a second level of shrinkage for any set of codes that may also be rarely used. Therefore, we exploit the hierarchical taxonomy of procedure codes and employ two stages of penalization. Note that such hierarchical increment can be generalized to introduce nested blocks by including one indicator for each (sub-)block level, for example, CPT codes 10000–69990 belong to a block denoting surgery. Within this block, there are several sub-blocks

denoting general surgery (10000–10022), integumentary system (10040–19499), and etc. Information for estimating coefficient of a certain block level comes from utilization of all codes and sub-blocks within this level.

The estimation of ridge regression is to minimize the negative log-likelihood plus a penalty function:

$$\begin{aligned} &P(\lambda_{\text{ridge}}, \omega_0, \omega_1, \omega_2) \\ &\quad = \lambda_{\text{ridge}}\big[\omega_0\big(\|\alpha_0\|_2^2 + \|\alpha_1\|_2^2 + \|\boldsymbol{\theta}\|_2^2\big) + \omega_1\big(\|\boldsymbol{\alpha_b}\|_2^2 + \|\boldsymbol{\gamma_{b1}}\|_2^2\big) \\ &\qquad + \omega_2\big(\|\boldsymbol{\eta_c}\|_2^2 + \|\boldsymbol{\zeta_{c1}}\|_2^2\big)\big]. \end{aligned}$$

The form of the penalties on the coefficients guides the properties of the model. The tuning parameter λ_{ridge} controls the strength of the penalty, and $\omega_0, \omega_1, \omega_2 \in [0, 1]$ allow varying penalties to different coefficients. In particular, $\lambda_{\text{ridge}}\omega_1$ controls variation in the block effects, and $\lambda_{\text{ridge}}\omega_2$ controls shrinkage of code effects toward their overall block effect. The shrinkage is particularly important for rare codes, which could yield extreme crude estimates.

A caveat is that we need to choose the value of λ_{ridge} and ω. Throughout this work, we introduce hierarchical shrinkage by fixing $\omega_0 = 0$, $\omega_1 = 0.5$, and $\omega_2 = 1$, which puts no penalty on the cohort, and restricts the penalization of block such that it is half as strong as the penalization of code. For λ_{ridge}, a sequence of 100 values is calculated corresponding to the regularization path. Because our primary goal with penalization is to simply stabilize rate ratios for codes with sparse data and/or zero counts, we choose the 95th λ_{ridge} along the sequence which gives a model with small to moderate penalization.

3.2. *Inference for ridge estimation in Poisson model.* Applying the ridge penalty has the effect of shrinking the estimates toward zero, which introduces bias but reduces the variance and stabilizes the estimates [Hoerl and Kennard (1970)]. Testing for potentially sparse and over-dispersed CPT code utilization data can benefit from this property as long as we could de-bias the estimate in order to perform valid inference. Here, we introduce a testing procedure for ridge regression in the Poisson family with log-link, referred to as the ridge test hereafter. We note that the proposed method can be adapted to any generalized linear model with a canonical link.

We simplify the notation and rewrite the model as

$$\log\big[E(\boldsymbol{Y}|\mathbf{X})\big] = \log(\boldsymbol{t}) + \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{Y}$ is the number of assignments for all patients and all codes, $\boldsymbol{t}$ is the offset denoting lengths of follow-up, $\mathbf{X}$ is the design matrix containing the intercept, indicator cohort_1, covariates $\mathbf{Z}$, indicators for all blocks: block_b, $\text{block}_b \cdot \text{cohort}_1$, and indicators for all codes: code_c, $\text{code}_c \cdot \text{cohort}_1$, and coefficient $\boldsymbol{\beta} = [\alpha_0, \alpha_1, \boldsymbol{\theta}^T, \boldsymbol{\alpha_b}^T, \boldsymbol{\gamma_{b1}}^T, \boldsymbol{\eta_c}^T, \boldsymbol{\zeta_{c1}}^T]^T$. The penalty function can be written as $P(\cdot) = \boldsymbol{\beta}^T \boldsymbol{\Lambda} \boldsymbol{\beta}$, where $\boldsymbol{\Lambda} = \lambda_{\text{ridge}} \cdot \boldsymbol{W}$, with penalty weight matrix $\boldsymbol{W} = \text{diag}\{\omega_0 \cdot \mathbf{1_0}, \omega_1 \cdot \mathbf{1_1}, \omega_2 \cdot \mathbf{1_2}\}$.

LEMMA 3.1. *We assume that we are in the common EMR situation where $n > p$ since the population sizes under study are commonly large and the number of observations in the Poisson regression is driven by both the number of patients and the dimension of the multivariate outcome. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the ridge shrinkage estimator with regularization parameter $\lambda_{\text{ridge}} > 0$. Then the de-biased estimator correcting for shrinkage bias is*

$$\hat{\boldsymbol{\beta}}_{\text{debias}} = \{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) + \boldsymbol{\Lambda}]^{-1}\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)\}^{-1}\hat{\boldsymbol{\beta}}_\lambda,$$

where $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda) = \mathbf{X}^T \operatorname{diag}(e^{\log(\boldsymbol{t})+\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda})\mathbf{X}/n$. Let $\boldsymbol{\beta}^*$ be the population coefficient satisfying model (2), and let $\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}$ be the population ridge coefficient from model (2) with penalty function $P(\cdot) = \boldsymbol{\beta}^T\boldsymbol{\Lambda}\boldsymbol{\beta}$ and $\boldsymbol{\Lambda} = \lambda_{\text{ridge}} \cdot \boldsymbol{W}$. Then we have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{debias} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, [\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}^*)[\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}})]^{-1}),$$

where $\boldsymbol{H}(\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}) = \mathrm{E}[\boldsymbol{X}\operatorname{diag}(e^{\log(t)+\boldsymbol{X}^T\boldsymbol{\beta}^*_{\boldsymbol{\Lambda}}})\boldsymbol{X}^T]$ and $\boldsymbol{\Omega}(\boldsymbol{\beta}^*) = \operatorname{Var}[\boldsymbol{X}(Y - e^{\log(t)+\boldsymbol{X}^T\boldsymbol{\beta}^*})]$.

A proof is detailed in Supplement B of the supplementary materials [Shi, Pashova and Heagerty (2017)]. Based on Lemma 3.1, a z-statistic for constructing p-value is

$$\big(\operatorname{diag}\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\text{debias}})[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\}\big)^{-\frac{1}{2}}\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{debias}},$$

where $\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\text{debias}}) = \mathbf{X}^T\{\operatorname{diag}[(\boldsymbol{Y} - e^{\log(\boldsymbol{t})+\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{debias}}})^2]\}\mathbf{X}$ is a sandwich estimator of the variance $\boldsymbol{\Omega}(\boldsymbol{\beta}^*)$. The confidence interval is

$$\hat{\boldsymbol{\beta}}_{\text{debias}} \pm Z_{1-\alpha/2}\big(\operatorname{diag}\{[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\hat{\boldsymbol{\Omega}}_n(\hat{\boldsymbol{\beta}}_{\text{debias}})[\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)]^{-1}\}\big)^{\frac{1}{2}}/\sqrt{n}.$$

Significance testing for regularized regression is a contemporary topic in the statistical literature, and there are very few publications addressing testing with ridge regression. Bühlmann (2013) developed a testing procedure for ridge regression in the high-dimensional setting in which the number of regression parameters p is larger than the sample size n, and assuming a deterministic design matrix. Similar to their work, we now account for shrinkage bias that results from penalization and derive the distribution of a de-biased estimate. However, we took a slightly different path: we de-biased by appropriately rescaling the penalized estimator instead of subtracting an estimated bias term. The later approach requires an initial consistent estimate of the true coefficient. In addition, our testing method tackles a problem that is different from the setting in Bühlmann (2013). Although the number of CPT codes is large, our problem remains essentially a low-dimensional problem because the number of observations in the Poisson regression is driven by both the number of patients and the dimension of the multivariate outcome for each patient, which is typically much larger than the number of procedure codes.

Therefore, information matrices such as $\boldsymbol{H}_n(\hat{\boldsymbol{\beta}}_\lambda)$ presented in Lemma 3.1 are not singular and can be inverted, and our estimator is not subject to the projection bias in high dimensions discussed in Bühlmann (2013). Our choice of ridge regression is purely for stabilizing estimates and not for solving the more common over-specification or singularity issue. Therefore, we can directly rescale the shrinkage estimator without the need to separately estimate a bias term.

3.3. *Patient-level clustering due to simultaneous analysis of multivariate outcomes*. With the use of ridge regression, we provide inference regarding the systematic associations with covariates such as healthcare system (site) for multiple CPT codes. Therefore, a vector of multivariate outcomes represents the data that is analyzed for each patient and characterizes the full set of CPT outcomes for that individual. Simultaneous regression with blocks of codes would then need to appropriately account for the multiple outcomes per patient. Standard GEE-type sandwich variance estimates are a simple way to account for the potential within-subject correlation for this situation [Bull (1998)].

3.4. *Provider-level clustering*. In Section 2.3, we discussed testing options when patients treated by the same provider are potentially correlated. In the same spirit, a sandwich variance estimator can be used to replace $\hat{\boldsymbol{\Omega}}_n$ discussed in Section 3.2 to generate a ridge test that accounts for provider-level clustering, which can be viewed as variations of a Generalized Estimating Equations (GEE) strategy [Diggle et al. (2002)]. Performance of the ridge test with and without a robust variance-covariance matrix is studied via simulation in Section 4.1.2.

4. Simulations. Previous research has compared group-wise association tests in the context of genome-wide association studies [see, for instance, Wu et al. (2011), Chapman and Whittaker (2008), Pan (2009), Basu and Pan (2011), Qi et al. (2015)], and relevant results are detailed in Supplement C of the supplementary materials [Shi, Pashova and Heagerty (2017)]. Generalizing these results to our context implies that burden tests may have increased power to detect association at the block level when the direction of code-specific effects are similar. In contrast, when code-specific effects may differ in direction, SKAT has been shown to be an effective testing strategy [Wu et al. (2011)]. When provider-level clustering is present, the L-KM method controls the type I error and increases power compared to competing methods [Qi et al. (2015)].

We focus our simulation studies on characterizing the finite sample operating characteristics of code-wise testing strategies where both common and rare codes are of interest. Sparse codes are likely to require exact methods to preserve the nominal type I error rate, while common codes are likely to require methods/models for overdispersed count data. We are not aware of any literature that provides a comprehensive characterization of test option performance in the CPT

code context, and such evaluation is necessary to provide recommendations for routine comparison of healthcare utilization.

For all of our simulation studies, we consider four classes of testing options:

- Analysis of the CPT count data using a LRT and a ET based on either the simple Poisson model or the more general negative binomial model.
- Analysis of the derived any/none binary indicator using Binomial methods: Fisher's ET and Binomial LRT.
- Simple two-sample t-test with unequal variances, or with a sandwich variance estimator (z-statistic) for correlated data, as a potential semiparametric method relying solely on moment assumptions.
- Ridge test that constructs p-value and confidence interval using de-biased estimator assuming either independent or correlated data.

We evaluate the performance in terms of type I error rate and power across a full range of rate parameters as might be encountered in healthcare utilization data. We also evaluate inference of the ridge regression in terms of type I error and coverage.

We consider three types of underlying data for our simulation studies:

- $Y_{si}^c \sim$ Poisson(rate $= \lambda_s^c$) with $\mathrm{E}[Y_{si}^c] = \mathrm{Var}[Y_{si}^c] = \lambda_s^c$. Note that $\Pr(Y_{si}^c = 0) = e^{-\lambda_s^c}$.
- $Y_{si}^c \sim$ negative binomial($\mu = \lambda_s^c, r = 2$) with $\mathrm{E}[Y_{si}^c] = \lambda_s^c$ and $\mathrm{Var}[Y_{si}^c] = \lambda_s^c(1 + \frac{1}{2}\lambda_s^c)$. Note that $\Pr(Y_{si}^c = 0) = (\frac{r}{\lambda_s^c + r})^r = (\frac{r}{\lambda_s^c + 2})^2$.
- $Y_{si}^c \sim$ zero-inflated Poisson($\pi^c = \frac{r}{1+r} = \frac{2}{3}, \lambda_s^c$) with $\mathrm{E}[Y_{si}^c] = \lambda_s^c$ and $\mathrm{Var}[Y_{si}^c] = \lambda_s^c(1 + \frac{1}{2}\lambda_s^c)$. There are two components: $Y_{si}^c = 0$ with probability $\pi^c = \frac{2}{3}$, and $Y_{si}^c \sim$ Poisson(rate $= \frac{\lambda_s^c}{1-\pi^c}$) with probability $1 - \pi^c$. Both generate zero-counts, so $\Pr(Y_{si}^c = 0) = \pi^c + (1 - \pi^c)e^{-\frac{\lambda_s^c}{1-\pi^c}} = \frac{2}{3} + \frac{1}{3}e^{-3\lambda_s^c}$.

We dichotomize the individual count data to create indicators, Z_{si}^c, analyzed using a Binomial LRT. Note that $\Pr(Y_{si}^c = 0) \to 0$ as $\lambda_s \to \infty$ in the negative binomial and Poisson model implying that for count data with a large mean, the dichotomized data would be all ones and, therefore, be uninformative for evaluation of rate differences. Finally, for ETs we aggregate individual data, Y_{si}^c and Z_{si}^c, into cohort-level totals Y_s^c and Z_s^c.

Our choice of generating models is to allow both standard Poisson models as well as alternative distributions that generate over-dispersed data. The primary role of the zero-inflated Poisson model is to allow evaluation of the flexibility or robustness of the negative binomial model for over-dispersed data that are outside the assumed class of models used for analysis. We parameterize the three data generating models such that they all have the same mean λ_s, which represents the average number of times a procedure was delivered to a patient during a fixed (one unit) follow-up time period. In addition, the negative binomial model and the zero-inflated Poisson model are chosen to have the same variance.

In addition to concerns of sparsity, over-dispersion, and model misspecification, we are also interested in the influence of provider-level clustering on validity of the tests, and the performance of testing strategies that account for correlated data. To this end, we assign each patient to a provider randomly with an average cluster size of three patients per provider, and introduce correlation between patients within provider by including a provider-level random variable. We consider two types of mean-preserving random variables that fluctuate the mean λ_s by provider p:

- $\lambda^{\beta}_{s,p} = \lambda_s \cdot \gamma^{\beta}_p$, where $\gamma^{\beta}_p \sim \text{Gamma}(\text{shape} = \frac{1}{\beta}, \text{scale} = \beta)$ is shared by patients within the same provider p, with $\text{E}[\gamma] = 1$, $\text{Var}[\gamma] = \beta$, and $\text{E}[\lambda^{\beta}_{s,p}] = \lambda_s$.
- $\lambda^{\sigma}_{s,p} = \exp[\log(\lambda_s) + b^{\sigma}_p]$, where $b^{\sigma}_p \sim \text{Normal}(0, \sigma^2)$ is shared by patients within the same provider p, with $\text{E}[\lambda^{\beta}_{s,p}] = \lambda_s \cdot e^{\frac{\sigma^2}{2}}$ but $\text{E}[\log(\lambda^{\beta}_{s,p})] = \log(\lambda_s)$.

4.1. *Code-wise test*: *Type I error rate*. In the following sections, we assess size of the testing options under imbalance sample sizes, sparsity, over-dispersion, model misspecification, and potentially correlated data introduced by provider behavior.

4.1.1. *Independent data*. In this section, we generate independent outcomes to estimate the type I error rate as the proportion of p-values less than the nominal α level of 0.05. We set $\log_{10}\lambda_0 = \log_{10}\lambda_1$ to range from -6 to 2, so that λ ($\equiv \lambda_0 = \lambda_1$) increases from 10^{-6} to 10^2 multiplicatively. Under this null, the rate ratio is one and both cohorts have equal variances. For each scenario considered, we conduct 5000 simulations with unequal samples sizes of $n_0 = 1000$ and $n_1 = 3000$ since this reflects the motivating example data.

The performance of each test varies by the average number of times a procedure was delivered in the sample, $(n_0 + n_1)\lambda$, which we refer to as the frequency. We will discuss selected results presented in Figure 1 by four regions of $\log_{10}\lambda$. They are region I: $[-6, -4)$, region II: $[-4, -2)$, region III: $[-2, -0.5)$ and region IV: $[-0.5, 2]$, which correspond to a frequency of less than 0.4, less than 40, between 40 and 1265, and over 1265, respectively, for the sample size of 4000. The comprehensive results with equal and unequal sample sizes are shown in Supplement D Figures 1 and 2 of the supplementary materials. We also evaluated coverage of the ridge test shown in Supplement D Figure 5 of the supplementary materials [Shi, Pashova and Heagerty (2017)].

For extremely low rates (region I), all methods have a type I error of nearly zero for any of the three types of simulated data. This is not surprising since a rare procedure that gets assigned to only one out of every 10,000 patients should provide little information unless sample sizes are extremely large.

When outcome rates are rare (region II), with an expected rate λ of 0.1 to 10 per 1000, or equivalently, a frequency of 0.4 to 40, we find that LRTs, t-test, and the ridge test have inflated type I error rates. Conversely, the three ETs still hold the type I error below the nominal level.

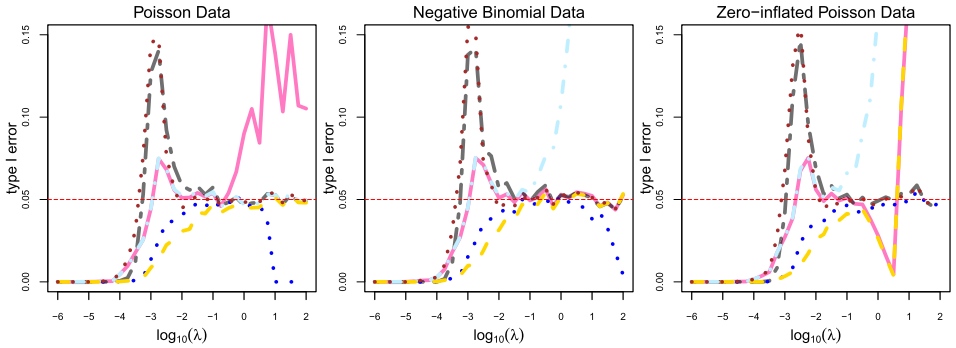


FIG. 1. *Type I error rates of CPT code-specific tests with unequal sample sizes* ($n_0 = 1000$, $n_1 = 3000$) *using Poisson data, negative binomial data, or zero-inflated Poisson data, each with a group mean of* λ *ranging from* 10^{-6} *to* 10^2, *plotted on* $\log_{10}$ *scale. Colored lines correspond to negative binomial LRT* (——); *negative binomial ET* (– – –); *Poisson LRT* (— · —); *Fisher's ET* (·····); *t-test* (– – –); *ridge test* (·····).

In region III when λ is approximately 100 per 1000, all of the tests perform well and have a type I error rate of around 0.05.

Finally, in region IV, the *t*-test and ridge test are the only two tests with a type I error rate near the nominal level for all three data-generating mechanisms. All of the model-based LRTs are subject to inflated type I error when the assumed model is incorrect. Specifically, the negative binomial LRT and ET break down when the true distribution is zero-inflated Poisson, although they are valid for both the distributions within their assumed class (negative binomial and Poisson). Also, when λ is large, the induced dichotomized data in the negative binomial and Poisson model are all ones and, therefore, the Binomial LRT and Fisher's ET will have a type I error rate of zero.

Our simulation results suggest that a valid and simple testing strategy for rate differences can be obtained from a dynamic test that uses the negative binomial ET if the total number of delivery in the sample (i.e., the frequency) is less than 40, and otherwise uses the semiparametric *t*-test. In order to evaluate the performance of such a procedure, we have calculated test size in additional simulations (in Supplement D Figure 6 of the supplementary materials [Shi, Pashova and Heagerty (2017)]), and find that the dynamic test tracks the conservative type I error of exact methods for low rates, but then enjoys robustness to model assumption for moderate and large rates. In additional simulations, we evaluated the threshold of 40 and find this appropriate with varying total sample size.

In summary, we find that no method can be reliably used across the entire spectrum of candidate rates that are encountered with CPT data. For rare rates exact testing methods are preferred, while for common rates robust methods such as the *t*-test and the ridge test perform well. Model-based count data LRT do not exhibit sufficient robustness to rare counts or model violation to be recommended for routine surveillance use.

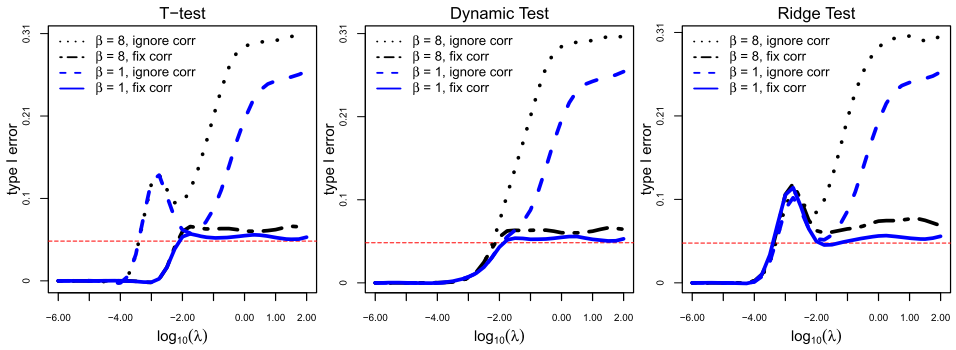


FIG. 2. *Type I error rates of t-test, dynamic test, and ridge test using negative binomial data with provider-level clustering. To introduce association, a mean-preserving random variable* $\gamma_p^\beta \sim \text{Gamma}(\text{shape} = \frac{1}{\beta}, \text{scale} = \beta)$ *with* $E[\gamma_p^\beta] = 1$ *and* $\text{Var}[\gamma_p^\beta] = \beta$ *is shared by patients treated by provider p. The cohorts have unequal sample sizes* ($n_0 = 1000$, $n_1 = 3000$), *each with a group mean of* λ *ranging from* 10^{-6} *to* 10^2, *plotted on* $\log_{10}$ *scale.*

4.1.2. *Correlated data.* In this section, we investigate influence of provider-level clustering, and evaluate performance of the proposed corrections in t-test, dynamic test, and ridge test that allow for correlation between patients within provider. Recall that we introduce correlation using two types of mean-preserving random variables (simulation settings in Section 4). First, for a fixed λ, we study how strength of correlation influences the type I error by setting variances of the random variables, defined by β and σ^2, to increase from $\frac{1}{16}$ to 4 multiplicatively. Second, we fix β and σ^2, and set λ to range from 10^{-6} to 10^2 to study sensitivity and robustness to correlated data under both rare and common code scenarios.

Figure 2 shows selected results which are type I error rates of t-test, dynamic test, and ridge test using negative binomial data with provider-level clustering. For tests that ignores the correlation between patients within provider, we observe two patterns: when CPT codes are rare, the testing procedures are not very sensitive to correlation among patients, and the type I error is not substantially inflated; when CPT codes are common, type I error exceeds the nominal value. In addition, the type I error increases when either the utilization rate λ increases or the correlation controlled by β and σ^2 gets stronger. We also see that when the covariance matrix is estimated accounting for correlation within provider clusters in each of the tests, the type I error is corrected to the nominal α level. Comprehensive results for Gamma and Normal random variables, and across different data generating distributions including negative binomial, Poisson, and zero-inflated Poisson are quite similar and are shown in Supplement D Figures 3 and 4 of the supplementary materials [Shi, Pashova and Heagerty (2017)].

4.2. *Code-wise test*: *Power.* To explore power, we focus on select event rates, $\lambda_0 = 1$ or 0.01, and empirically evaluate power as a function of various rate ratios,

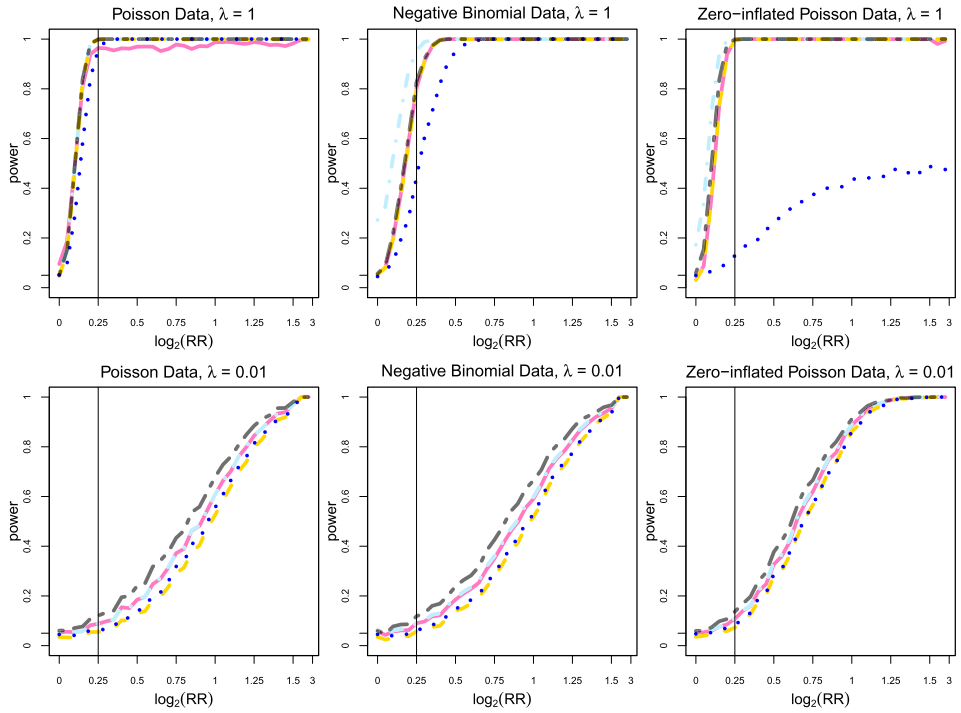


FIG. 3. *Power of CPT code-specific tests with unequal sample sizes* ($n_0 = 1000$, $n_1 = 3000$) *using negative binomial data, Poisson data, or zero-inflated Poisson data, each with a mean of* $\lambda_0 = 1$ *or* 0.01 *in cohort* 0, *and a rate ratio on* $\log 2$ *scale* ($\log_2 \frac{\lambda_1}{\lambda_0}$) *ranging from* 0 *to* 1.5. *Colored lines correspond to negative binomial LRT* (——); *negative binomial ET* (– –); *Poisson LRT* (·—·—·); *Fisher's ET* (·····); *T-test* (– –).

$\lambda_1 = \text{RR} \cdot \lambda_0$. We let $\log_2 \text{RR}$ range from 0 to 3, so that the RR increases from 1 to 8 multiplicatively, and we conduct 2000 simulation replications for each situation. Figure 3 shows selected results for unequal sample sizes similar to our motivating data, and show a monotone increase in power with increasing RR, and a small loss of power for exact and robust methods relative to the correctly specified LRTs. Comprehensive results for equal and unequal sample sizes are quite similar and are shown in Supplement D Figures 7 and 8 of the supplementary materials [Shi, Pashova and Heagerty (2017)]. A few observations are notable, with the first being that application of Fisher's ET when the event rates are large may result in a substantial loss of power. When $\lambda_0 = 1$, we see that most test procedures achieve power greater than 80% for $\log_2 \text{RR} > 0.25$ for all three data-generating mechanisms, while Fisher's ET has power $< 50\%$ for all RR values under a zero-inflated Poisson model. However, with lower event rates (e.g., $\lambda_0 = 0.01$) we see a small reduction in power with use of ETs. Therefore, the power plots reinforce recommendations based on preservation of test size: ETs appear valid and reasonably

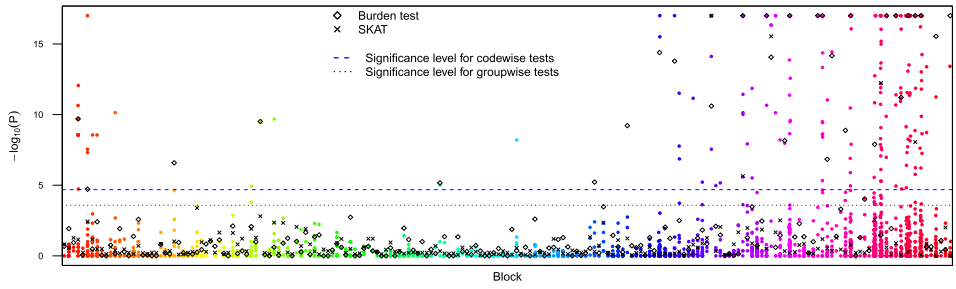
powered for low event rates; while robust methods are valid and retain power for common event rates.

5. Application: Comparing healthcare utilization between Henry ford health system and kaiser permanente. The development and evaluation of statistical methods for comparing rates of medical procedure utilization is motivated by the Back pain Outcomes using Longitudinal Data (BOLD) project, which enrolled 5239 patients aged 65 years and older with a new episode of back pain [Jarvik (2012)]. In order to enroll more than 5000 patients, recruitment was conducted from three healthcare systems. Primary scientific questions focus on medical interventions such as early radiologic imaging and subsequent patient reported pain and function outcomes. In order to combine EMR data across the three systems, we need to understand any differences in procedure endorsement across the sites. Therefore, we focus on CPT coding data across all domains including imaging, laboratory, and diagnostic procedures. Specific sub-cohorts can be defined using the demographic or clinical information. Here, we focus on the cohorts defined by the enrollment site for a patient since both geographic and healthcare system differences may be associated with different CPT coding patterns. We compare healthcare utilization between two largest sites: Henry Ford Health System in Detroit and Kaiser Permanente in Northern California, which include 4040 patients.

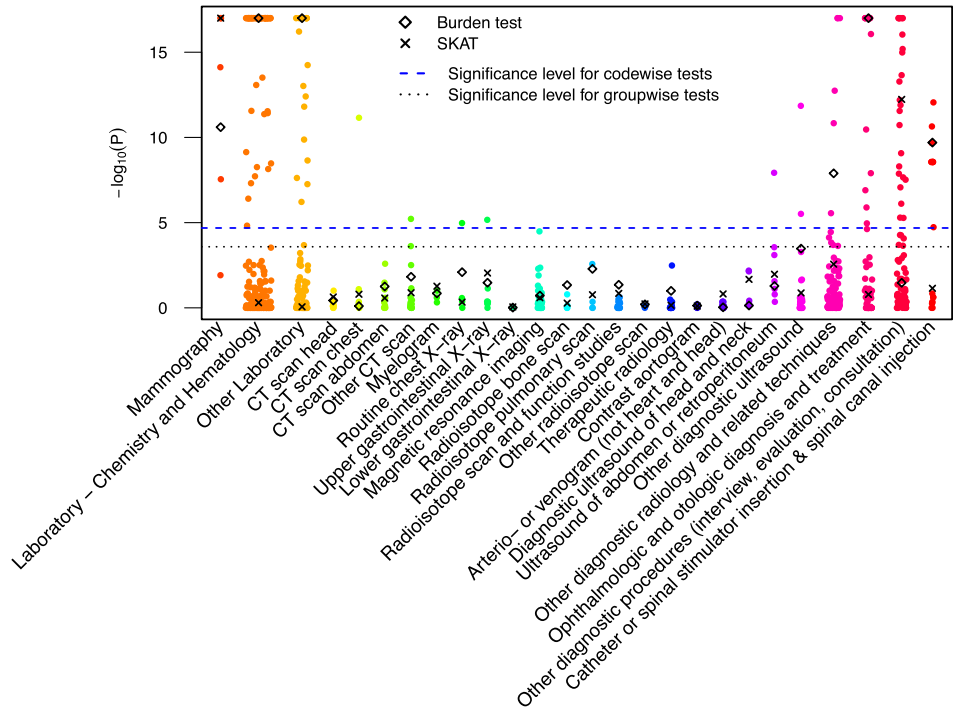
5.1. *Hypothesis testing*. First, we investigate the significance of the difference in code utilization for individual codes and blocks of codes defined by the CCS-Services and Procedures. We use the dynamic test for both code-specific comparison and the burden test; for block-based inference, we use both the burden test and SKAT. We present code-wise p-values on the $-\log_{10}$ scale for all CPT codes in a "Manhattan plot" (Figure 4), for which the codes in a block are contiguous and plotted with the same color. We truncate any p-value at 10^{-17}. We also add the group-wise p-values to the Manhattan plot, one for each block. We include two horizontal lines which are the Bonferroni corrected significance thresholds for code-level and block-level comparisons. There are a total of $C = 2424$ CPT codes with nonzero counts, and based on the CCS classification we have $B = 192$ blocks.

Figure 4(a) shows that there are many codes with utilization differences that are statistically significant, and that these codes tend to cluster in select domains. We detected significant difference among 31 out of 192 blocks using the burden test, and only 5 using the SKAT. Specifically, in 27 blocks, the burden test rejects the null hypothesis while SKAT does not, and there is only 1 block for which the SKAT has a significant result in contrast to the result from the burden test.

We zoom in on select blocks to investigate detailed patterns as shown in Figure 4(b). The burden test for "Laboratory—Chemistry and Hematology" rejected the null hypothesis, while the SKAT p-value did not reach the block-wise significance level. We find that this is driven by the abundance of codes whose utilizations



(a)



(b)

FIG. 4. *A full Manhattan plot for code-wise comparison of Henry Ford Health System and Kaiser Permanente plotted by block, overlaid with results from the group-wise comparison using the Burden test and SKAT for each block. The y-axis is truncated at a p-value of* 10^{-17}. *Bonferroni corrected significance levels for code-wise and group-wise tests are shown. Panel* (b) *is a zoom in version of the Manhattan plot for select blocks.*

are consistently higher at one site, a situation in which the burden test has higher power.

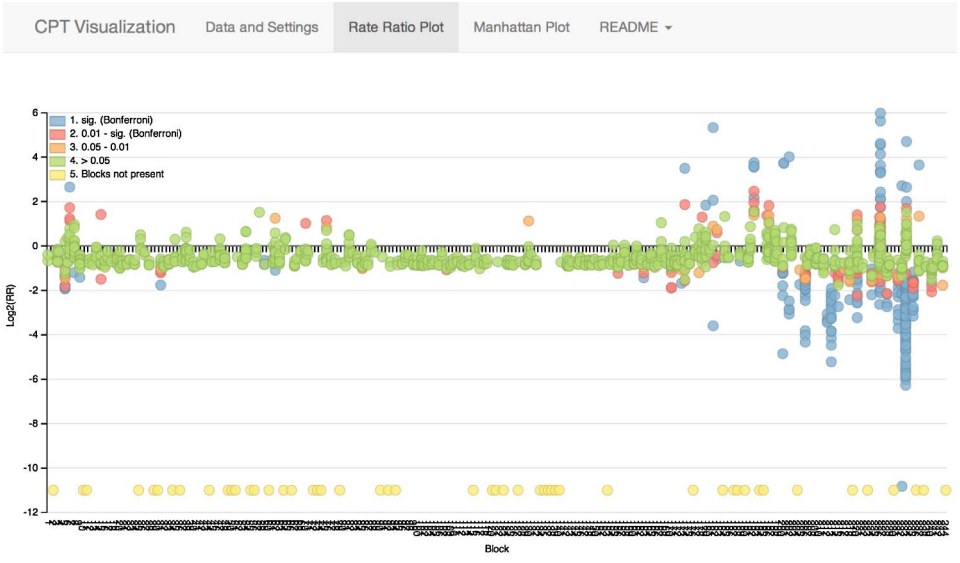
In the block called "Mammography", although both the burden test and SKAT give a significant result, SKAT gives a much smaller p-value than the burden test.

Looking into the data, we see that Henry Ford Health System uses exclusively so-called G-codes for recording of mammography, while Kaiser Permanente uses the common numeric codes for mammography. We learned that G-codes refer to digital mammography for screening or diagnosis, whereas numeric codes refer to nondigital (film) mammography. Therefore, if the cost of breast cancer screening was the focus of analysis, then minor differences in codes may be important to capture. On the other hand, if overall interest is in the rate of patient screening regardless of imaging technology, then G-codes and numeric codes may be combined to define a general mammography procedure and the detailed differences will not be important. Such distinctions relate to the choice of code-specific or block-specific inference procedures that we present, as well as the trade-off between sensitivity and power as illustrated by comparing the results from the burden test and from SKAT.

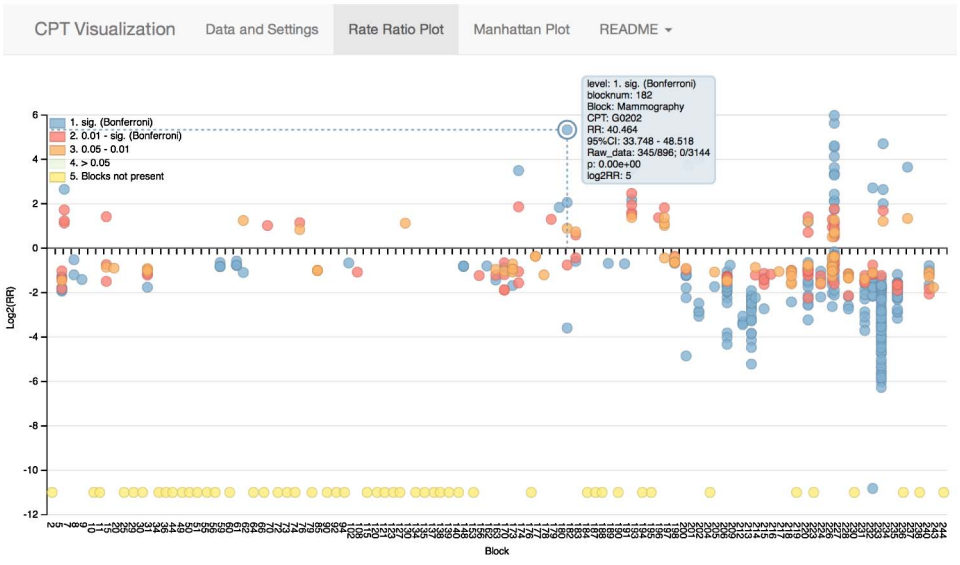
Our methods are intended to identify specific codes, or groups of codes, that appear to have different recorded utilization. Additional investigation is required to separate whether the finding corresponds to actual differences in patient care, or whether coding variation through use of alternative codes may explain differences in observed endorsement rates. In our use of these methods with our healthcare delivery system studies, we have used the signals from our testing procedure to engage in discussion with the individual systems to ultimately attribute observed differences to practice or coding variation.

5.2. *Rate ratio estimation and inference.* A key aspect of understanding differences in CPT endorsement is the direction and magnitude of rate differences. Therefore, we also estimate rate ratios for all codes simultaneously to compare the utilization pattern in Henry Ford Health System to Kaiser Permanente, adjusting for age category, sex, and race. Due to potential sparse codes, we use penalized Poisson regression as detailed in Section 3, and we provide inference for ridge regression using methods detailed in Section 3.2. In the BOLD study, patients were recruited through primary care clinics, and information on their primary care provider is available. To illustrate the practical use of our proposed methods that account for provider-level clustering discussed in Section 3.4, we adopt the primary care provider as our level of clustering which is appropriate for the study design. In the BOLD study, we have 4040 patients and 1819 providers.

To display the point estimation results, we use dynamic graphical methods that plot the estimated code-specific rate ratios on a $\log_2$ scale versus the block to which each code belongs to (Figure 5). We color-code each point according to the significance of the code from the hypothesis test using the ridge test detailed in Section 3.2. The p-values are split into four regions, each corresponding to a color. They are: $(0, \alpha]$, $(\alpha, 0.01]$, $(0.01, 0.05]$, and $(0.05, 1]$, where α is the Bonferroni corrected significance level. The summary plot displays the healthcare utilization pattern over the entire spectrum of all possible procedures, but detail on individual codes can be revealed using dynamic graphical methods, and as shown in Figure 5,



(a)



(b)

FIG. 5. *Rate ratio estimates comparing Henry Ford and Kaiser Permanente for each CPT code based on a penalized Poisson regression with ridge penalty. Code-specific rate ratios* ($\log_2$ *scale*) *are plotted against the block that the each code belongs to, color-coded according to four levels of* p*-values*: $(0, \alpha]$, $(\alpha, 0.01]$, $(0.01, 0.05]$, *and* $(0.05, 1]$, *where* α *is the Bonferroni corrected significance level. The plot function can dynamically provide additional information for each point showing the block, the code, the rate ratio, the* 95% *confidence interval, the* p*-value, and the raw data, as illustrated with one point in panel* (b).

we display both statistical information and medical details for any given CPT code. For blocks that are not present in the data, we plot place holders at the bottom using a unique color. We note that individual CPT code rate ratios are shrunk to a block level rate ratio due to our choice of penalization that incorporates the block-code hierarchical structure.

There are a variety of factors that may drive measured utilization differences, including healthcare practices, coding regulations, data quality issues, and differences in patient characteristics. Our methods can reveal differences that should be followed up with additional investigation into the underlying drivers for the observed differences. To this exploratory end, we make the plot interactive which displays the tool-tip detailing the information of the block, the code, the rate ratio, the 95% confidence interval, the p-value, as well as the raw data. One can also filter on select results based on p-value categories. Pointing to a specific point in the plot, for example, the G0202 in the "Mammography" block, we can see from the raw data that Henry Ford uses the G codes exclusively, while Kaiser Permanente uses the five-digits codes, as is discussed in Section 5.1. Such pattern drives the estimated rate ratio to be high and the p-value to be low.

The summary rate ratio plot serves as an interpretable tool for clinical researchers and data managers to explore healthcare utilization patterns among subcohorts. We have used this tool as part of our data quality control, and for providing potential alternative explanations that need to be considered in comparative utilization analyses across observed patient subgroups. We implement the interactive plot in both an R package and a shiny application. The shiny app is available online at https://xu-rita-shi.shinyapps.io/CPT_visualization/.

6. Discussion. Contemporary biomedical research is now leveraging the electronic medical records for both comparison of alternative treatment options and to generate individual predictions. Increasingly, there are large networks of hospitals or healthcare systems that are assembled to provide sizable cohorts. With these efforts comes the need to compare patterns of utilization across sub-groups within modern cohorts, either to understand systematic issues with respect to data quality or coding variation, or to compare utilization across patient subgroups defined by treatment or medical indication. Therefore, we have developed multilevel hypothesis testing and rate ratio estimation methods that can be used either for evaluation of potential data issues or for comparative inference.

First, we detailed statistical testing methods for evaluating differences in procedure assignments between two groups, and provide inference at both the code and block levels. To compare utilization at the code level, we discussed the potential likelihood ratio test and conditional exact tests. We focused on three candidate distributions: the Poisson, negative binomial, and Binomial distribution if data are dichotomized. When comparing rare procedure codes which might lead to low power and violation of assumptions for the asymptotic χ^2 approximation,

the conditional exact test provides a viable option. We also considered semiparametric testing using the two-sample t-test. We learned from our simulation study that different tests work well in different scenarios, and the dynamic test tracks the conservative type I error of exact methods for low rates, but then enjoys robustness to model assumption for moderate and large rates. To compare utilization on block level, we transferred methods from genome-wide association studies to the EMR context, including the burden test and the sequence kernel association test. Both the burden test and SKAT evaluate utilization patterns by combining a block of similar codes, which may substantially increase power for rare codes in particular.

Second, we detail estimation and inference of utilization rate ratios via penalized Poisson regression with a tailored form of penalty that takes advantage of the hierarchical structure of the CPT codes. Our proposed method shrinks the code-specific estimates to the block level, effectively borrowing information from all other codes within the same block. Such shrinkage is especially important for rare codes for which individual rate ratio estimates may be highly unstable. We also develop inference methods that account for shrinkage bias and construct statistical tests (p-values) and confidence intervals using the distribution of a de-biased estimate.

Third, we consider provider behavior as an important driver of patterns in healthcare utilization and expanded the inference method to account for potential correlation within provider. We learned from simulation that for rare CPT codes, testing is not sensitive to provider clustering, because there is not much information to be influenced by within-provider correlation. In contrast, correlation does inflate the type I error among common outcomes, and the amount of inflation increases with the strength of the correlation and the mean of the outcome. We are also able to control the inflated type I error under correlated data back to its nominal value by correcting the variance-covariance estimate.

Finally, we ultimately provide interpretable dynamic graphical tools that can help researchers to explore and interpret the healthcare utilization patterns. We use a CPT code version of the genomic Manhattan plot to display testing results, and we use an interactive plot to present both the significance and the magnitude of rate differences. The interactive plot enables us to see useful global information and select detailed information that facilitates discovery of key utilization differences. Although we focus on CPT coding differences, the general testing and estimation framework is also applicable to other forms of structured EMR data such as diagnostic coding data (ICD-9 or ICD-10) and is particularly useful when any coding system can be hierarchically organized.

A potential limitation of our work is the need to further investigate whether any statistical finding corresponds to actual differences in patient care, or whether coding variation through use of alternative codes may explain differences in observed endorsement rates. Another limitation is that in our data application, we only had primary care provider IDs and did not have detailed specialty care provider information to illustrate how one could use an extension of the GEE-type variance

estimator to accommodate nonnested clustering. In addition, although we consider accounting for patient-specific follow-up time to partially account for missing data, tailored methods targeting missing data is an important future direction in the use of EMR data. Last, the longitudinal nature of EHR data has great potential for research to understand the temporal changes in patient treatment history and patient health status, which requires future work.

SUPPLEMENTARY MATERIAL

Supplement A: Comprehensive discussion on code-wise two-sample testing options (DOI: 10.1214/17-AOAS1028SUPPA; .pdf). We provide detailed review of testing strategies that are candidates for the evaluation of variation in code endorsement rates across cohorts.

Supplement B: Proof of Lemma 3.1 (DOI: 10.1214/17-AOAS1028SUPPB; .pdf). We provide a proof of Lemma 3.1.

Supplement C: Comprehensive review of simulation results comparing group-wise association tests (DOI: 10.1214/17-AOAS1028SUPPC; .pdf). We provide a review of relevant results in previous research comparing group-wise association tests.

Supplement D: Comprehensive plots of type I error and power (DOI: 10.1214/17-AOAS1028SUPPD; .pdf). We provide additional supporting plots that show the type I error and power of all tests with equal/unequal sample sizes using generated data of independent observations or under provider-level clustering.

REFERENCES

BASU, S. and PAN, W. (2011). Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology* **35** 606–619.

BEHRMAN, R. E., BENNER, J. S., BROWN, J. S., MCCLELLAN, M., WOODCOCK, J. and PLATT, R. (2011). Developing the Sentinel System—A national resource for evidence development. *N. Engl. J. Med.* **364** 498–499.

BENTLEY, P. N., WILSON, A. G., DERWIN, M. E., SCODELLARO, R. and JACKSON, R. E. (2002). Reliability of assigning correct current procedural terminology—4 E/M codes. *Ann. Emerg. Med.* **40** 269–274.

BÜHLMANN, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19** 1212–1242. MR3102549

BULL, S. B. (1998). Regression models for multiple outcomes in large epidemiologic studies. *Stat. Med.* **17** 2179–2197.

CHAPMAN, J. and WHITTAKER, J. (2008). Analysis of multiple SNPs in a candidate gene or region. *Genetic Epidemiology* **32** 560–566.

COLLINS, F. S., HUDSON, K. L., BRIGGS, J. P. and LAUER, M. S. (2014). PCORnet: Turning a dream into reality. *Journal of the American Medical Informatics Association* **4** 576–577.

DIGGLE, P. J., HEAGERTY, P. J., LIANG, K.-Y. and ZEGER, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. *Oxford Statistical Science Series* **25**. Oxford Univ. Press, Oxford. MR2049007

HOERL, A. and KENNARD, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

HOLT, J., WARSY, A. and WRIGHT, P. (2010). Medical decision making: Guide to improved CPT coding. *South. Med. J.* **103** 316–322.

JARVIK, J. G. (2012). Study protocol: The back pain outcomes using longitudinal data (BOLD) registry. *BMC Musculoskelet Disord.* **13** 64.

KING, M. S., LIPSKY, M. S. and SHARP, L. (2002). Expert agreement in Current Procedural Terminology evaluation and management coding. *Arch. Intern. Med.* **162** 316–320.

KING, M. S., SHARP, L. and LIPSKY, M. S. (2001). Accuracy of CPT evaluation and management coding by family physicians. *J. Am. Board Fam. Pract.* **14** 184–192.

LEE, S., EMOND, M. J., BASHED, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., NHLBI GO EXOME SEQUENCING PROJECT—ESP LUNG PROJECT TEAM, CHRISTIANI, D. C., WURZEL, M. M. and LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91** 224–237.

MADSEN, B. E. and BROWNING, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5** e1000384.

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2004). Marginal modeling of multilevel binary data with time-varying covariates. *Biostatistics* **5** 381–398.

MIGLIORETTI, D. L. and HEAGERTY, P. J. (2007). Marginal modeling of nonnested multilevel data using standard software. *Am. J. Epidemiol.* **165** 453–463.

MORGENTHALER, S. and THILLY, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res.* **615** 28–56.

MORRIS, A. P. and ZEGGINI, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology* **34** 188–193.

PAN, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology* **33** 487–507.

PRZYBOROWSKI, J. and WILENSKI, H. (1940). Homogeneity of results in testing samples from Poisson series with an application to testing clover seed for dodder. *Biometrika* **31** 313–323. MR0002070

QI, Y., WEEKS, D. E., TIWARI, H. K., YI, N., ZHANG, K., GAO, G., LIN, W., LOU, X., CHEN, W. and LIU, W. (2015). Rare-variant kernel machine test for longitudinal data from population and family samples. *Hum. Hered.* **80** 126–138.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, G. K. (2010). EdgeR: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 515–524.

SHI, X., PASHOVA, H. and HEAGERTY, P. J. (2017). Supplement to "Comparing healthcare utilization patterns via global differences in the endorsement of current procedural terminology codes." DOI:10.1214/17-AOAS1028SUPPA, DOI:10.1214/17-AOAS1028SUPPB, DOI:10.1214/17-AOAS1028SUPPC, DOI:10.1214/17-AOAS1028SUPPD.

WU, M. C., LEE, S., CAI, T., LI, Y., BOEHNKE, M. and LIN, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89** 82–93.

X. SHI
P. J. HEAGERTY
DEPARTMENT OF BIOSTATISTICS
UNIVERSITY OF WASHINGTON
SEATTLE, WASHINGTON 98195-7232
USA
E-MAIL: xushi@uw.edu
heagerty@uw.edu

H. PASHOVA
AXIO RESEARCH
2601 4TH AVENUE, SUITE 200
SEATTLE, WASHINGTON 98121
USA
E-MAIL: ninap@axioresearch.com