SUPPLEMENTARY MATERIAL

Supplement A: Comprehensive Discussion on Code-wise Two-sample Testing Options

(doi: [10.1214/17-AOAS1028SUPPA](https://doi.org/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. 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. We review standard asymptotic likelihood ratio tests (LRTs) and conditional exact tests (ETs) for count and binary data with a goal of characterizing the applied options.

1. Count Outcome Asymptotic and Exact Tests: Simple Poisson Model for Rates. For count data, a natural model is a Poisson distribution. Specifically, the counts for code assignment are assumed to be characterized by a rate parameter, λ_s^c , with $Y_{si}^c \sim \text{Poisson}(\lambda_s^c)$, $i = 1, \dots, n_s$.

In this situation, we may wish to test whether code endorsement varies by cohort, i.e., $H_0: \lambda_0^c = \lambda_1^c$. 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 Poisson assumption was valid, then the 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 $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c \sim \text{Poisson}(n_s \lambda_s^c)$, and apply the Poisson ET which does not depend on large sample approximations [Przyborowski and Wilenski (1940)]. The Poisson ET computes the p -value based on the fact that, under the null, $Y_1^c \mid Y_0^c + Y_1^c \sim \text{Binomial}(n = Y_0^c + Y_1^c, p = \frac{n_1}{n_0 + n_1})$.

2. Count Outcome Asymptotic and Exact Tests: Negative Binomial Model for Overdispersion. In cohorts where there are both relatively healthy and extremely ill patients, there will be overdispersion in the CPT utilization counts, 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: $Y_{si}^c \sim \text{negative binomial}(p = p_s^c = \frac{\lambda_s^c}{\lambda_s^c + \frac{1}{\phi^c}}, r = 1/\phi^c)$, $i = 1, \dots, n_s$, where ϕ^c is the code-specific overdispersion parameter which is shared by patients across cohorts. The mean and vari-

ance are parameterized as $E[Y_{si}^c] = \lambda_s^c$ and $\text{Var}[Y_{si}^c] = \lambda_s^c(1 + \lambda_s^c\phi^c)$. Note that when $\phi^c = 0$, the model reduces to the Poisson model.

The sum of negative binomial variables with the same p is still negative binomial, and since the patient-level count data for code c within a cohort s are i.i.d. with the same p_s^c , the cohort-level data summary counts are also negative binomial: $Y_s^c = \sum_{i=1}^{n_s} Y_{si}^c \sim \text{negative binomial}(p' = \frac{n_s \lambda_s^c}{n_s \lambda_s^c + \frac{n_s}{\phi^c}}, r' = \frac{n_s}{\phi^c})$. Note that the probability $p' = p$, while the mean and variance are $E[Y_s^c] = n_s \lambda_s^c$ and $\text{Var}[Y_s^c] = n_s \lambda_s^c(1 + \lambda_s^c\phi^c)$.

We take parallel asymptotic and exact testing strategies as with the Poisson model. When the LRT χ^2 approximation can be assumed to be valid, we use the negative binomial LRT with patient-level data Y_{si}^c . Note that the overdispersion parameter will inflate the standard errors, making the test results more conservative than the Poisson LRT. When the expected cell counts are too small, we can instead rely on the negative binomial ET to calculate the conditional probability using the cohort-level data Y_s^c [Robinson, McCarthy and Smyth (2010)]. For a given code, we wish to test whether the CPT code assignment varies by cohort, i.e., $H_0: \lambda_0^c = \lambda_1^c \equiv \lambda^c$ without the requirement of asymptotic methods. Under the null, we have $p_0^c = p_1^c = \frac{\lambda^c}{\lambda^c + \frac{1}{\phi^c}}$ and $Y_s^c \sim \text{negative binomial}(p' = \frac{\lambda^c}{\lambda^c + \frac{1}{\phi^c}}, r' = \frac{n_s}{\phi^c})$. Thus $Y_0^c + Y_1^c \sim \text{negative binomial}(p'' = \frac{\lambda^c}{\lambda^c + \frac{1}{\phi^c}}, r'' = \frac{n_0 + n_1}{\phi^c})$. We compute the p -value via the conditional probability

$$\Pr(Y_0^c = y_0^c, Y_1^c = y_1^c \mid Y_0^c + Y_1^c = y_0^c + y_1^c) = \binom{y_0^c + \frac{n_0}{\phi^c} - 1}{y_0^c} \cdot \binom{y_1^c + \frac{n_1}{\phi^c} - 1}{y_1^c} / \binom{y_0^c + y_1^c + \frac{n_0 + n_1}{\phi^c} - 1}{y_0^c + y_1^c}.$$

Although the conditional probability eliminates the common rate parameter, λ^c , p -value computation still require the common overdispersion parameter ϕ^c . We can either vary the value of the nuisance parameter over a range of plausible values, or we can estimate ϕ^c based on the patient-level data Y_{si}^c using maximum likelihood. Specifically, we use negative binomial regression methods that allow separate group rates (e.g., do not impose null for λ_0^c and λ_1^c) and estimate the common overdispersion parameter. Because the test statistic has an exact distribution, with large sample sizes we expect substituting a consistent estimator of the nuisance parameter to yield similar performance as if the true nuisance parameter were known, and the sample size of EMR data is often sufficiently large. However, if the sample size is truly insufficient, estimation of the overdispersion parameter with small sample sizes has been studied intensively for differential expression detection in RNA-seq data [Robinson and Smyth (2008); Wu, Wang and Wu (2012)],

which can serve as alternative methods. In our simulation studies we evaluate the performance of the negative binomial ET with a plug in dispersion parameter.

The negative binomial distribution is a natural model for overdispersed count data that characterized unobserved heterogeneity through use of a continuous latent variable. A key advantage of the negative binomial is a generalized variance function which may allow valid testing of rate ratios. However, in certain situations the population under study may be comprised of a mixture of discrete subgroups such that a continuous latent variable model does not capture the true heterogeneity. When discrete heterogeneity is present, latent class or mixture models may describe the population distribution for count data, but often will require additional parameters that specify the mixture proportions and the subgroup count distributions. Therefore, the use of mixture models for testing will typically lead to a model that under the null will constrain multiple parameters which leads to a primary test statistic that has multiple degrees of freedom rather than a single focused test statistic for rate ratios. Power may be poor unless mixture model alternatives are carefully chosen such as fixing the mixture probabilities, or constraining feature of the component mixtures. For example, zero-inflated Poisson models are a special case of a discrete mixture, but have one component of the mixture specified as a point mass at zero. We focus our evaluation on the use of standard overdispersed count data models and their associated tests, but recognize that additional exploration of testing options using more complex mixture models is a direction of important future work. However, we do evaluate the performance of simple Poisson and negative binomial tests when data may in fact be generated from a more complex mixture mechanism such as a zero-inflated Poisson.

3. Count Outcome Asymptotic and Exact Tests: Semiparametric Two-sample t -test. 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. However, with large sample sizes and any underlying distribution, we can use the two-sample t -test as a semi-parametric method for testing. Use of the t -test for inference with CPT codes requires understanding the performance of the method for both common and rare codes, and below we focus on three central issues.

First, with a large overall sample and equal groups sizes, the t -test is known to be robust to both inequality of variance and non-normality, and maintains a valid test size due to the central limit theorem [Boneau (1960)].

Second, with comparison groups that have unequal variances, Welch’s t -test would be preferred as it correctly estimates the standard deviation in the denominator of the t statistic. In addition, Welch’s t -test has been shown to perform better than the Student’s t -test (assuming equal variances) when the data are normally distributed [Ruxton (2006); Zimmerman and Zumbo (1993)].

Finally, while the Welch’s t -test naturally accommodates unequal variances, it does not necessarily eliminate the effect of non-normality. Although EMR data may provide large sample sizes, when the outcome rate is rare, the sample mean of discrete counts may not be approximately normally distributed. In particular, the effect of non-normality lies in the higher order moments (skewness and kurtosis) that arise in the expansion of the mean and variance of the t statistic. It has been shown that the kurtosis has little effect on the asymptotic distribution of the t statistic, and has no effect when the sample sizes are equal. The skewness does affect the asymptotic mean and variance of the t statistic. Thus, comparing the sample means for two populations with different shapes requires caution and perhaps necessitates alternative testing strategies [Miller Jr (1997)]. Ultimately, t -tests provide simple asymptotic inference, but the performance in terms of size and power for EMR data should be investigated to guide applied statistical practice for both rare and common outcomes.

4. Any/None Outcome Asymptotic and Exact Tests. Use of CPT counts and associated rates of endorsement 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 potential count data as a derived binary outcome since the clinical significance is indicated by any endorsement of the code.

Statistically the methods for testing a binary outcome are well characterized. An any/none outcome indicates whether a patient was ever assigned a code during his or her visits over the year. These outcomes can be modeled as $Z_{si}^c \sim \text{Bernoulli}(p_s^c)$ for a patient, and $Z_s^c = \sum_{i=1}^{n_s} Z_{si}^c \sim \text{Binomial}(n_s, p_s^c)$ 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. Under the null, the conditional distri-

bution of the cell counts is hypergeometric, with conditional probability

$$\Pr(Z_0^c = z_0^c, Z_1^c = z_1^c \mid Z_0^c + Z_1^c = z_0^c + z_1^c) = \binom{n_0}{z_0^c} \cdot \binom{n_1}{z_1^c} / \binom{n_0 + n_1}{z_0^c + z_1^c}.$$

In summary, key practical issues include: whether to adopt asymptotic tests or exact tests; whether to consider a count or indicator outcome; and whether or how to account for overdispersion in CPT counts.

References.

- BONEAU, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological bulletin* **49**.
- MILLER JR, R. G. (1997). *Beyond ANOVA: basics of applied statistics*. CRC Press.
- PRZYBOROWSKI, J. and WILENSKI, H. (1940). Homogeneity of results in testing samples from Poisson series. *Biometrika* **31**.
- 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**.
- ROBINSON, M. D. and SMYTH, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9** 321-332.
- RUXTON, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*. **17** 688-690.
- WU, H., WANG, C. and WU, Z. (2012). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics* **14** 232-243.
- ZIMMERMAN, D. W. and ZUMBO, B. D. (1993). Rank transformations and the power of the Student t test and Welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*. **47** 523-539.