Detecting divergent health care providers

i2i

May 2, 2017

1 Introduction

Many analyses done by i2i involve detecting divergent health care providers[1], which are divergent in the sense that they perform a particular 'event' more often than health care providers of the same type (their "peer group"). In such an analysis, every patient undergoes a Bernoulli trial: they either get a particular treatment (event) or they don't (no event). An event might be for example staying in the hospital overnight, being operated on, having two diagnoses in parallel, etc.

Currently, we have a few different methods for detecting statistically significant divergence of providers, and a proposal for a better method to replace all of our current methods. This document describes this proposal and is intended for an external statistician to help answer the question: "Is our proposed new method, a justifiable method?".

2 Proposed method

Our method consists of 2 parts. First, we run a hypothesis test, using the null hypothesis that the event probability is the same at each provider within a peer group, after accounting for certain covariates. For each provider this leads to a certain margin above which it will be labeled as divergent. Secondly, we estimate the between provider standard deviation that is not accounted for by the covariates we use in the first step. This is then added to the margin, and the amount by which a provider exceeds this margin, is converted into an equivalent cost in euros.

2.1 Hypothesis test

Our test first requires classification of patients into different buckets based on their features, such as age and diagnosis, because part of the variation between providers can be explained by differences in their patient populations. For example, older patients might require a particular expensive treatment more

[1] by health care provider we mean an institution, such as a hospital or a pharmacy.

often than younger patients, in which case we do not want to unfairly label a provider as divergent simply because it has many older patients. We create a different bucket for every distinct combination of values of each covariate (for example, a bucket might contain male patients between 30 and 40 years old, having diagnosis X).

In the following, j labels health care providers and i labels different patient buckets. Then n_{ij} is the number of patients and o_{ij} the number of observed treatments or, more generally, events. With a single index these symbols represent the total per provider: $o_{\bullet j} = \sum_i o_{ij}$, $n_{\bullet j} = \sum_i n_{ij}$.

In our model, $o_{ij} \sim \text{binom}(n_{ij}, p_{ij})$, where the o_{ij} are observed random variables and the n_{ij} and the p_{ij} are known and unknown parameters, respectively. Since we benchmark providers against a peer group, we are interested in the average treatment probability over all providers in the peer group

$$\rho_i = \frac{\sum_j n_{ij} p_{ij}}{\sum_j n_{ij}},$$

with maximum likelihood estimate

$$\hat{\rho}_i = \frac{\sum_j o_{ij}}{\sum_j n_{ij}}.$$

Now we wish to test if a provider treats its patients more often than its peergroup, taking differences in patient population into account, but without testing every patient type seperately. Therefore, as the null hypothesis for provider j we choose

$$H_0 : \frac{\sum_i n_{ij} p_{ij}}{n_{\bullet j}} = \frac{\sum_i n_{ij} \rho_i}{n_{\bullet j}},$$

with the alternative hypothesis

$$H_1 : \frac{\sum_i n_{ij} p_{ij}}{n_{\bullet j}} > \frac{\sum_i n_{ij} \rho_i}{n_{\bullet j}}.$$

So the null hypothesis says that the total fraction of patients treated is consistent with the assumption that every patient would get the same treatment in every hospital in the peer group.

The corresponding test statistic is $o_{\bullet j}/n_{\bullet j}$ (the fraction of patients treated) which, under the null hypothesis, is a sum of binomial random variables, with expectation

$$e_j = \frac{\sum_i n_{ij} \rho_i}{n_{\bullet j}}$$

and variance

$$\sigma_j = \frac{\sqrt{\sum_i n_{ij} \rho_i (1 - \rho_i)}}{n_{\bullet j}},$$

where we have assumed the o_{ij} to be independent, so we may sum their variances. We estimate e_j and σ_j as $\hat{e}_j = \frac{\sum_i n_{ij} \hat{\rho}_i}{n_{\bullet j}}$ and $\hat{\sigma}_j = \frac{\sqrt{\sum_i n_{ij} \hat{\rho}_i (1 - \hat{\rho}_i)}}{n_{\bullet j}}$

We reject the null hypothesis when

$$\frac{o_{\bullet j}}{n_{\bullet j}} > \hat{e}_j + t * \hat{\sigma}_j.$$

The value t is the threshold parameter, that determines the balance between sensitivity and specificity. Assuming $n_{ij} \cdot p_{ij}$ is large enough, we may apporximate the o_{ij}, and hence $o_{\bullet j}$, as Gaussian, in which case a value of $t = 2$ should lead to a significance level of $\alpha = 2.5\%$.

The procedure described here is extended in the next section, where we calculate an additional parameter τ which is combined with σ_j.

As a side note, care should be taken in classifying the patients to prevent having too many buckets in which many providers have $n_{ij} = 0$. In that case, providers having $n_{ij} = 0$ essentially don't contribute to ρ_i. In an extreme case, there may be only one provider k having patients in bucket i (so $n_{ij} = 0$ when $j \neq k$) in which case $\rho_i = p_{ik}$ and the expectation for that provider in that bucket is based entirely on the provider itself, and there is no benchmarking based on the peer group.

2.2 Two level model with between provider standard deviation

In the previous model any provider that significantly deviated from its expected value based on its peergroup, was considered divergent. This is problematic because adjusting for the patient characteristics available to us is suspected to be insufficient to effectively take into account all"legitimate" sources of variation between providers. Furthermore, any tiny deviation will be significant in this model, as long as the provider is large enough.

Therefore we turn the model from the previous section into a two-level model, where the e_j are no longer parameters, but are themselves drawn from a distribution, to model variation in the true effect among providers (a "random effects" model):

$$e_j \sim \mathcal{N}(\bar{e}, \tau^2)$$

$$\frac{o_{\bullet j}}{n_{\bullet j}} \sim \mathcal{N}(e_j, \sigma_j^2)$$

so that

$$\frac{o_{\bullet j}}{n_{\bullet j}} \sim \mathcal{N}(\bar{e}, \sigma_j^2 + \tau^2).$$

This is similar to the model described in section 3.1 of [1]. In this model e_j and σ_j are estimated as before, and the null hypothesis will be rejected when

$$\frac{o_{\bullet j}}{n_{\bullet j}} > \hat{e}_j + t * \sqrt{\hat{s}_j^2 + \hat{\tau}^2}.$$

The task is then to find a good estimate of τ, the true "between provider standard deviation". For this we follow [2], explaining the intermediate steps.

In the notation of [2], we would have

$$T_j = \frac{o_{\bullet j}}{n_{\bullet j}} - \hat{e}_j$$

and weights

$$w_j = \frac{1}{n_j}$$

and observed weighted mean

$$\bar{T} = \frac{\sum_j w_j T_j}{\sum_j w_j}.$$

Here T_j is the deviation of the observed fraction of patients treated from the fraction that is expected based on the model in the previous section. Under the null hypothesis of that model, the expected value of T_j would be 0 for every provider.

To estimate τ^2, we take the observed sample variance in T and subtract what the sample variance in T would be under the the hypothesis that all provider effects were drawn from the same distribution (i.e. $\tau = 0$). This is then divided by the bias in the sample variance.

We can write the observed sample variance as $\frac{Q}{V_1}$, where

$$Q = \sum_{j=1}^{k} w_i (T_j - \bar{T})^2,$$

$V_1 = \sum_{i=1}^{k} w_i$ is the sum of the weights as in [3] and k is the total number of providers in the peer group. For the expected sample variance, $\frac{Q_0}{V_1}$, we have:

$$Q_0 = E\left[\sum_{j=1}^{k} w_j (T_j - \bar{T})^2\right]$$

$$= \sum_{j=1}^{k} w_j E[(T_j - \bar{T})^2]$$

because the weights are non-random. Under the aforementioned hypothesis that $\tau = 0$, we have

$$E[(T_j - \bar{T})^2] = \frac{k-1}{k}\sigma_j^2,$$

where σ_j^2 is the variance in T_j [2] and $\frac{k-1}{k}$ is the sample bias. This leads to [3]

$$Q_0 = \sum_{j=1}^{k} w_j \frac{k-1}{k} \sigma_j^2.$$

As explained in the appendix, the weighted sample variance has a bias of $1 - \frac{V_2}{V_1^2}$ (where $V_2 = \sum_j w_j^2$), so our final estimate of τ is

$$\hat{\tau}^2 = \frac{Q - \sum_{j=1}^{k} w_j \frac{k-1}{k} \hat{\sigma}_j^2}{C},$$

where

$$C = V_1 - \frac{V_2}{V_1}.$$

If $\hat{\tau}^2$ calculated this way becomes negative, we set it to 0.

2.3 Interpretation of outcome

The amount by which a provider has to change in order to no longer be divergent is reported to the provider as $(\frac{o_{\bullet j}}{n_{\bullet j}} - (e_{\bullet j} + t \cdot \sqrt{\hat{\sigma}_j^2 + \hat{\tau}^2})) * n_{\bullet j}$. This amount is multiplied by the estimated average cost of each excess event, to report the deviation in terms of money. In our reports this is refered to as "risico-omzet", which is calculated for a range of indicators at the level of (usually) either specialty or diagnosis (group).

In this way we do not have the binary outcome of either rejecting or accepting the null hypothesis, but we are sensitive to the actual effect size. The calculated "risico-omzet" actually depends both on the true effect size and on the size of the provider (the same effect size, could be significant in a large provider while it is not significant in a smaller provider).

3 Appendix: Bias in weighted sample variance

Suppose we have a sample T_j, $j = 1, 2, ..., k$ and (non-random) weights w_j, $j = 1, 2, ..., k$. The T_j are independent and identically distributed. The weighted sample mean is

$$\bar{T} = \frac{\sum_j w_j T_j}{V_1} \tag{1}$$

[2] Here we approximate the variance in $\hat{e}_j$ as 0, because it is based on the entire peer group and is thus expected to vary much less than $\frac{o_{\bullet j}}{n_{\bullet j}}$, which is the fraction of treated patients in a single hospital. This might lead to an underestimation of Q_0 and hence an overstimation of τ and hence to some false negatives.

[3] Using weights $w_j = 1/\sigma_j^2$ as in [2] leads to a more elegant equation. However, because in the calculation of $\hat{e}_j$ we use the size of the providers as weights, it seems more consistent to do so here as well.

where $V_1 = \sum_j w_j$. Let the weighted true population mean be μ. Then the expected weighted sample variance s is

$$\begin{aligned} E[s] &= \frac{E\left[\sum_j w_j (T_j - \bar{T})^2\right]}{V_1} \\ &= \frac{E\left[\sum_{j=1}^{k} w_j ((T_j - \mu) - (\bar{T} - \mu))^2\right]}{V_1} \\ &= \frac{E\left[\sum_{j=1}^{k} w_j ((T_j - \mu)^2 + (\bar{T} - \mu)^2 - 2(T_j - \mu)(\bar{T} - \mu))\right]}{V_1} \\ &= \frac{E\left[\sum_{j=1}^{k} w_j (T_j - \mu)^2\right]}{V_1} - E\left[(\bar{T} - \mu)^2\right] \\ &= \operatorname{Var} T - \operatorname{Var} \bar{T} \end{aligned} \tag{2}$$

where $\operatorname{Var}(X)$ represents the true weighted population variance in X.

We can write $\operatorname{Var}(\bar{T})$ as

$$\begin{aligned} \operatorname{Var}(\bar{T}) &= \operatorname{Var}\left(\frac{\sum_j w_j T_j}{V_1}\right) \\ &= \frac{1}{V_1^2} \sum_j \operatorname{Var}(w_j T_j) \\ &= \operatorname{Var}(T) \frac{\sum_j w_j^2}{V_1^2} \\ &= \operatorname{Var}(T) \frac{V_2}{V_1^2} \end{aligned} \tag{3}$$

Substituting equation 3 back into equation 2, we have

$$E[s] = \operatorname{Var}(T)\left(1 - \frac{V_2}{V_1^2}\right), \tag{4}$$

where $(1 - \frac{V_2}{V_1^2})$ is the sample bias.

References

[1] Haley E. Jones and David J. Spiegelhalter. The identification of unusual health-care providers from a hierarchical model. *American Statistical Association*, 65(3), 2011.

[2] M. Borenstein, L. Hedges, and H. Rothstein. Meta-analysis: Fixed effect vs. random effects, pages 11–16. `"http://www.meta-analysis.com/downloads/Meta-analysis%20fixed%20effect%20vs%20random%20effects.pdf"`.

[3] Weighted arithmetic mean. "https://en.wikipedia.org/wiki/Weighted_arithmetic_mean#Weighted_sample_variance".