

The use of count data models in biomedical informatics evaluation research

Jing Du,¹ Young-Taek Park,² Nawanan Theera-Ampornpunt,² Jeffrey S McCullough,¹ Stuart M Speedie²

¹Division of Health Policy and Management, University of Minnesota, Minneapolis, Minnesota, USA

²Institute for Health Informatics, University of Minnesota, Minneapolis, Minnesota, USA

Correspondence to

Jing Du, Division of Health Policy and Management, 420 Delaware Street SE, Minneapolis, MN 55455, USA; dux031@umn.edu

Received 17 March 2011

Accepted 5 June 2011

Published Online First

29 June 2011

ABSTRACT

Objectives Studies on the impact and value of health information technology (HIT) have often focused on outcome measures that are counts of such things as hospital admissions or the number of laboratory tests per patient. These measures with their highly skewed distributions (high frequency of 0s and 1s) are more appropriately analyzed with count data models than the much more frequently used variations of ordinary least squares (OLS). Use of a statistical procedure that does not properly fit the distribution of the data can result in significant findings being overlooked. The objective of this paper is to encourage greater use of count data models by demonstrating their utility with an example based on the authors' current work.

Target audience Researchers conducting impact and outcome studies related to HIT.

Scope We review and discuss count data models and illustrate their value in comparison to OLS using an example from a study of the impact of an electronic health record (EHR) on laboratory test orders. The best count data model reveals significant relationships that OLS does not detect. We conclude that comprehensive model checking is highly recommended to identify the most appropriate analytic model when the dependent variable being examined contains count data. This strategy can lead to more valid and precise findings in HIT evaluation studies.

INTRODUCTION

Count data consist of non-negative integers that represent the number of times a discrete event is observed.¹ In many health information technology (HIT) research contexts the dependent variable of interest fits this definition. Typical examples of count data are the number of office visits, hospital admissions, adverse drug events (ADEs), laboratory tests, and rates of cardiac arrest. This type of data presents a number of analytic challenges including: (a) a large and perhaps disproportionate number of zero values; (b) a relatively high frequency of small integer values; and (c) non-constant variance (where the variance of the residuals differs for different ranges of independent variables).

In recent years, statisticians, econometricians, and social science researchers have focused attention on modeling count data due to observed problems with commonly used analytic methods. Ordinary least squares (OLS) regression is often inappropriate because count data violate the underlying assumptions of OLS regression: normality and constant variance. This can result in

inaccurate estimates of standard errors, and misleading p values and consequent confidence intervals. It can also produce estimates of dependent variable values that are accurate for one subsample but inaccurate in other subsamples from the same population.²

One suggested approach is to apply OLS to the logarithm of a count variable to transform it into a more normal distribution.³ However, there are several drawbacks of such transformations including: (a) zero values are not taken into account⁴; (b) predicted meaningless negative values for the dependent variable are possible; (c) normality of the estimated values may not be achieved; (d) parameter estimates may be uninterpretable; and (e) retransformation may result in less accurate point estimates.² Thus, it is problematic to use OLS to analyze count data.

BRIEF REVIEW OF COUNT DATA MODELS

More effective alternatives have been developed to better account for the characteristics of count data. Regrettably, many researchers still rely on OLS to analyze count data and relatively few biomedical informaticians make use of count data models to examine the impact of HIT on such data. Our search of the literature in three leading research databases (PubMed, Academic Search Premier, and Elsevier ScienceDirect) revealed that only a limited number of reviewed papers used a count data model in the past 5 years, with little justification presented for their model selection decision.^{5–10} This is consistent with a recent review on the use of statistical analysis in the biomedical informatics literature. Descriptive and elementary statistics were most frequently used in the *Journal of the American Medical Informatics Association* and the *International Journal of Medical Informatics*, but few regression techniques for count data were in evidence.¹¹

This paper highlights the need for greater use of count data models, including Poisson regression (PR), negative binomial regression (NBR), hurdle regression (HR), zero-inflated Poisson regression (ZIPR), and zero-inflated negative binomial regression (ZINBR). The following discussion attempts to explain these data analysis methods for biomedical informaticians so that they may make use of them in their research for better and more precise statistical analyses. It attempts to avoid using statistical and econometric theories and terminologies wherever possible. For further details, the reader is referred to the work of Cameron and Trivedi, which covers both statistical and econometric analysis of count data models.¹ In the

following we cite biomedical informatics literature employing the discussed techniques when such articles exist.

Poisson regression

As the benchmark model for count data, the Poisson distribution models the probability of a number of events occurring within a given time, distance, or volume interval. It shares many similarities with OLS, but the residuals (differences between actual and predicted values) are assumed to follow the Poisson distribution instead of the normal distribution. Unlike OLS, which transforms the dependent variable to normalize the residuals, the log transformation in the Poisson distribution guarantees that the predicted values of the dependent variable will be zero or positive. The use of PR is illustrated by a study of the impact of point-of-care personal digital assistant (PDA) use on resident documentation that found no significant decrease in the number of documentation discrepancies due to PDA use.¹⁰

Negative binomial regression

When over-dispersion (the magnitude of the variance exceeds the magnitude of the mean) occurs, an NBR model is more appropriate. Since an NBR become a PR as over-dispersion declines, the main difference between a PR and an NBR is their variances. The consequence of over-dispersion in PR is underestimation of standard errors, which tend to be larger and more accurate in NBR. The fact that NBR converges to PR makes it possible to carry out a model comparison between them.¹² Empirically, NBR gives more accurate estimates than PR in most cases.¹³ An example of applying NBR because of over-dispersion in the number of ADEs was a study that examined the influence of a medication profiling program on the incidence of those ADEs.⁷ The authors concluded that a medication profiling program added to an e-prescribing system did not reduce ADEs.

Hurdle regression

Over-dispersion can occur in counts of events when the initial event and later events are generated by different processes. For example, an individual patient may initiate an office visit when he/she is concerned about certain symptoms. However, once the patient has an initial encounter, the number of subsequent visits is often determined by the physician. The hurdle model is designed to deal with count data generated from such systematically different processes. It assumes that zero and positive count values for the dependent variable come from different data generating mechanisms and as a result there may be more zero count cases than could be accounted for by a single generating mechanism. A hurdle model analysis consists of two parts. The first part models the probability of a count of zero for a case using a logit or probit regression. If the dependent variable has a positive value, the 'hurdle is crossed' and the second part models the positive count group using PR or NBR.¹⁴

Zero-inflated regression

Zero-inflated regression (ZIR) models also assume that a high frequency of zero counts (excessive zeros) may be due to more than one underlying process. Consider the case where the dependent variable of interest is the number of medications prescribed for patients as evidenced by insurance claims during a specified period. Some will have no medications recorded because they lack prescription drug insurance coverage, while others with such coverage will have a zero medication count because there is no medical need. Still other patients may both have insurance coverage and prescribed medications, thus having a medication count of one or more. Patients without insurance

are certain to have a count of zero medications. They are referred to as a certain zero group (CZG) and the probability of being in the CZG is modeled in the first part using a logistic regression. Patients in the latter two scenarios are considered to come from the not certain zero group (NCZG) and are modeled by PR (in ZIPR) or NBR (in ZINBR) in the second part that allows for the possibility of a predicted zero number of medications. The use of a ZINBR model is illustrated in a recent investigation of the impact of health information exchange (HIE) on emergency department (ED) utilization. It found that greater HIE information access was associated with increased ED visits and inpatient hospitalizations among patients with diabetes and asthma.⁵

Zero-inflated regression is similar to HR in many aspects. First, it is designed to accommodate the excessive zeros problem that can occur in count data. Second, it is a composite model that consists of two parts. The primary difference between ZIR and HR is that the logit part of HR models the probability that a patient comes from the zero count group. Unlike HR, the first part of ZIR differentiates the source of zero counts and thus predicts the probability that a patient is from the CZG using a logit model. Then either PR or NBR is applied in the second part to model the patients in the NCZG. The major difference between ZIPR and ZINBR is the same as that between PR and NBR, which means that ZINBR can better handle over-dispersion of count data.

AN APPLICATION OF COUNT DATA MODELS

To demonstrate the utility of count data models in comparison to OLS regression, we analyzed a dataset from a study we conducted examining the impact of an electronic health record (EHR) on the number of laboratory tests ordered in an ED. The six models used were OLS, PR, NBR, HR, ZIPR, and ZINBR. Details of this study are described elsewhere.¹⁵ The dependent variable was the number of laboratory tests ordered during an asthma patient's first encounter (index visit) at one of three EDs. The independent variables included the patient's demographic data (age, gender), if the patient had prior clinical information in the EHR at the time of the visit (Yes or No), the patient's burden of illness as measured by the Charlson comorbidity index, and site of care. Figure 1 is a frequency distribution that reveals a large number of patients with no laboratory test orders and much smaller numbers of patients with counts greater than zero. It is evident that the laboratory test data are not normally distributed and are highly concentrated in a few small values,

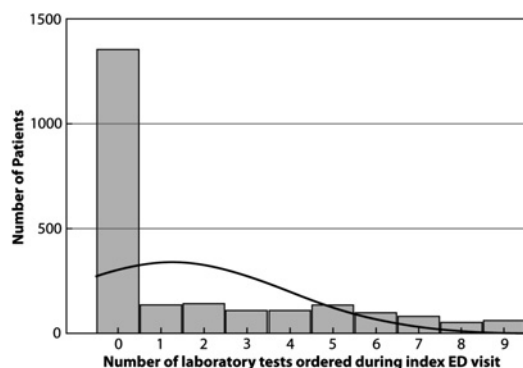


Figure 1 Observed frequency distribution of the number of laboratory tests ordered during index emergency department (ED) visits for asthma patients.

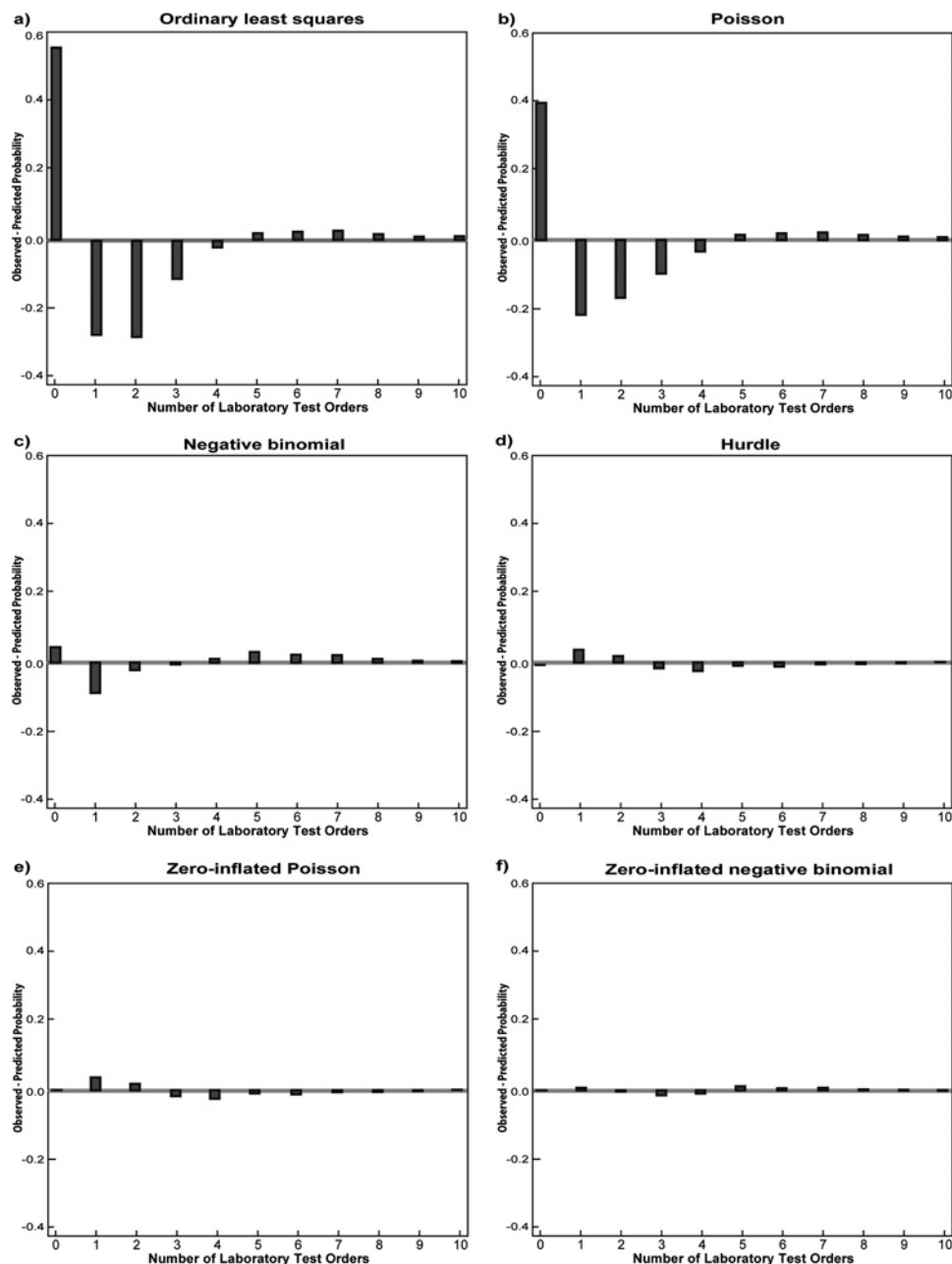
which makes intuitive sense given the characteristics of healthcare utilization. Most asthma patients have no laboratory tests during their index visit because many of their problems may be diagnosed by physical examination and patient interview; only a few patients may require further laboratory testing.

Selecting the appropriate model

A less formal way to evaluate the fit of a count data model is to plot the differences between the observed probabilities and the predicted probabilities of each count value (number of laboratory tests ordered) for each model. That probability is defined as the proportion of cases that take on a particular integer value over the range of such values contained in the data. Figure 2 is a series of plots of the differences between observed and predicted probabilities across all of the models this paper addresses. It clearly shows that both OLS (A) and PR (B) greatly underestimate the probability of a zero count and overestimate the probability of ordering one, two, three or four laboratory tests, indicating their poor fit to the count data. Negative

binomial regression (C) shows a notable improvement over PR in that it is better at predicting the probability of zero. This can be attributed to the fact that NBR explicitly allows for the observed over-dispersion of count data ($\text{variance}=11.5 > \text{mean}=2.1$). However, NBR still overestimates the probability of ordering one, two or three laboratory tests and underestimates the probabilities of ordering four or more laboratory tests, suggesting that an improvement in prediction is still possible. Compared with PR and NBR, both HR (D) and ZIPR (E) are better at predicting zero counts and generate predicted probabilities that are closer to those observed. However, ZIPR slightly outperforms HR in terms of predicting zero counts since it takes into account the source of zero counts. ZINBR (F) shows a moderate improvement over ZIPR and HR. It is not only better at predicting zero counts, but also does better in prediction than other models for the rest of the counts, although it slightly overestimates the probabilities of ordering three and four laboratory tests. Overall, ZINBR appears to most accurately fit the observed distribution of probabilities.

Figure 2 A graphical comparison of the fit of ordinary least squares and count data models to the number of observed laboratory test orders.



The more formal tests used to assess the fit of count data models include the likelihood ratio (LR) test, Akaike's information criterion (AIC), and the Bayesian information criterion (BIC).^{16–18} Before using PR, an over-dispersion check should always be performed. If the distribution of the count data is found to violate PR's equidispersion (equality of variance and mean) assumption (as in the example data), NBR should be performed instead. In practice, both SAS and Stata automatically compute a dispersion parameter (α) in an NBR analysis to assess whether the mean and variance significantly differ from each other.^{12 19} If α equals zero, a PR analysis is preferred. If α is significantly greater than zero, the count data are over-dispersed and are better estimated using an NBR. In our case, α is highly significant (data not shown) and indicates that NBR is preferred over PR.

The LR test can be used to compare nested models (ie, when one model is an extension of the other). In this set, PR is nested in NBR, PR is nested in HR, and ZIPR is nested in ZINBR. Based on the log likelihood values in table 1, comparing PR with NBR yields an LR test statistic of 4284 ($=-2(\log \text{likelihood for PR}-\log \text{likelihood for NBR})=-2(-6238-(-4096))$) with a p value of <0.001 . When PR is compared with HR, the LR test statistic is 4422. Comparing ZIPR with ZINBR yields an LR test statistic of 402 (both $p>0.001$). From this we can conclude that NBR is preferred over PR, HR is preferred over PR, and ZINBR is preferred over ZIPR.

In practice, the information criteria such as AIC and BIC are used to help determine the appropriateness of alternative models, especially when they are not nested. The lower the value of the AIC or BIC, the better the model. In table 1, PR, NBR, HR, ZIPR, and ZINBR are ranked according to their AIC and BIC values, with ZINBR preferred due to it having the lowest AIC and BIC. For those requiring a more comprehensive test, the Vuong test has been adapted by Greene to compare non-nested models such as ZIPR versus PR, and ZINBR versus NBR.^{20 21}

Interpreting the results

Given all of the comparisons presented, it is evident that the ZINBR model is the preferred one for this data. It demonstrates the best visual fit in plotting the differences between the actual and predicted probabilities. The AIC, BIC, and LR test all indicate that the ZINBR model is the best fit for the data.

Estimated coefficients from the OLS and ZINBR models are presented in table 2 for comparison purposes. Although both the OLS and the preferred count data model identify age, comorbidity ratings, and site of care as significant contributors to the number of laboratory tests ordered, gender and EHR were revealed to be additional significant predictors by the better fitting ZINBR count data model.

In order to explain the large number of zero counts, ZINBR estimates the coefficients of the independent variables assuming

two different processes. One process determines whether or not a patient will have any tests (count of zero) and thus addresses membership in the CZG (first or logit part). Such a process might well be represented by a group of asthma patients who are making use of the ED for primary care and simply require a new prescription for their inhaler. The second process determines the number of laboratory tests, including zero, that were ordered for the patient (second or negative binomial part). The underlying assumption is that this process might result in no laboratory tests but for a different reason than in the CZG. For example, some patients may be adequately assessed and treated solely by a physical examination and require no laboratory testing.

Electronic health record, age, gender, and comorbidity are all statistically significant predictors of membership in the group that is certain to have zero laboratory tests ordered during the index visit. The EHR result is interpreted as follows. When holding all other variables constant, the existence of a patient's clinical information in the EHR was associated with a decrease in the odds of membership of the CZG by a factor of 0.49 ($OR=e^{(-0.705)}=0.49$, 95% CI 0.34 to 0.72). In more qualitative terms, patients without EHR information had a higher expectation than similar patients with EHR information of experiencing a care process that did not include any laboratory test orders. Similarly, the gender result is interpreted as compared with females, being male increases the odds of membership in the CZG by a factor of 1.32 ($OR=e^{0.275}=1.32$, 95% CI 1.08 to 1.61) when holding all other variables in the model constant. In other words, male patients had a greater expectation than females of experiencing a care process that did not include any laboratory test orders. It is important to note that neither was a significant predictor in the OLS analysis.

The second part (negative binomial component) models the process of determining how many laboratory tests are ordered for a patient if he/she has a chance to have any tests ordered during their index visit. In table 2, the site2 coefficient is interpreted as follows. Holding all the other variables constant, patients who have the opportunity to have a laboratory test ordered at site2 during their index visit would have a 45% ($RR=e^{0.373}=1.45$, 95% CI 1.13 to 1.86) higher mean number of laboratory tests than such asthma patients at site3. A similar result was observed for site1. It should be noted that the same independent variable may be significant in one part of the count data model but not in the other part.

We conclude from these results of the ZINBR model that asthma patients with information in an EHR have approximately twice the odds of having laboratory tests ordered for them during their ED visit. However, this information is not predictive of how many tests are ordered.

DISCUSSION

This article examined how count data models (PR, NBR, HR, ZIP, and NBR models) can be applied to analyze problems in

Table 1 Model comparison test statistics

Test statistic	Model*					
	OLS	Poisson	Negative binomial	Hurdle	Zero-inflated Poisson	Zero-inflated negative binomial
Log likelihood	-6021	-6238	-4096	-4027	-4027	-3826
Akaike's information criterion	12060	12494	8211	8091	8091	7690
Bayesian information criterion	12112	12546	8269	8194	8194	7799

***Bold** indicates the preferred model.
OLS, ordinary least squares.

Table 2 Estimated coefficients for ordinary least squares (OLS) and zero-inflated negative binomial regression (ZINBR) in predicting the number of laboratory test orders

Variable	Models*		
	OLS	Zero-inflated	
		Logit	Negative binomial
Constant	-1.612	2.159	0.753
Adjustors			
Age	0.052	-0.029	0.008
Gender (ref group=female)	-0.117	0.275	0.064
Comorbidity	1.106	-0.765	0.106
Site1† (ref group=site3)	0.581	-0.041	0.354
Site2† (ref group=site3)	0.343	0.342	0.373
Indicators of EHR			
EHR (ref group=no EHR)	0.440	-0.705	0.028
Site1×EHR	0.041	0.433	0.001
Site2×EHR	-0.535	0.754	-0.025

*Italics indicate $p \leq 0.05$ for the coefficient, **bold** indicates the preferred model.

†Site1 and site2 represent the dummy variables for the three sites from which data were obtained.

EHR, electronic health record.

HIT evaluations. The utility of these modeling strategies for count data with large numbers of zeros and over-dispersion has been illustrated through an example that examines factors affecting the number of laboratory test orders in an ED encounter. The ZINBR analysis, in contrast to the OLS analysis, revealed that an asthma patient without an EHR record had a greater expectation than a patient with an EHR record of experiencing a care process that did not include any laboratory test orders.

This is consistent with the Mekhjian *et al* report that the average number of laboratory orders per patient per diagnosis related group (DRG) increased by up to 50% in the 2 years after the introduction of a computerized physician order entry system.²² Our results are also consistent with Vest's investigation of the impact of HIE on ED utilization, where ZINBR was employed in demonstrating that greater HIE information access was also associated with greater utilization of services as represented by ED visits and inpatient hospitalizations among patients with asthma and diabetes.⁵

It is recognized that the finding reported in the example is at variance with much of the literature on EHR impact, which has reported reductions in the numbers of laboratory tests.^{23–32} It is unlikely that the contrary finding is due to the use of the count data analysis. Rather it is most likely explained by the focus on a small subset of patients with asthma. Our work with congestive heart failure or diabetes patients in the same project from which this example was drawn is in greater agreement with the finding of a reduction in laboratory test orders when an EHR is present.¹⁵ One possible explanation for our finding may well be that the differential effect was observed because the impact of clinical data in an EHR may vary by disease state.

Our study demonstrates, for our data, the conceptual and statistical advantages of these fairly recent statistical estimation methods over the traditional OLS estimation approach. The example highlighted the value of count data models in uncovering previously hidden findings and providing additional insight into the factors that affect test volumes. Although no single count data model works best for all situations, the family of models provides the potential for a better statistical fit to observed data than OLS. The illustrative example shows that OLS analysis presents an incomplete set of findings most likely

due to violations of its underlying assumptions of normality and constant variance. In the example, although both OLS and ZINBR found age, comorbidity ratings, and site of care to be significant contributors to test volumes, the ZINBR model uniquely identified the additional significant influences of gender and the existence of EHRs on test ordering.

As an example of how count data models might well reveal statistically significant predictors that were otherwise undetected, consider the Bates *et al* study of the impact of the computerized display of charges on the test volumes.³³ The study reported that although there was a lower rate of test ordering in the group associated with a computerized display of charges, it was not statistically significant. It is worth noting that their conclusions were apparently based on multiple linear regression analyses using OLS that were adjusted for age, gender, race, primary insurer, and DRG weight. Given that the number of laboratory test orders here is likely to exhibit the characteristics of count data, there is reason to believe that count data models would be a better fit to the data and thus more appropriate than multiple linear regression in quantifying the impact of EHR on test ordering. If count data models had been applied, the less accurate estimates of standard errors or misleading p values, if any, might have been avoided and a significant impact for the intervention might have emerged. We are not arguing that the intervention did, in fact, have such an effect but only that since there was a numeric difference in favor of the intervention, a count data analysis would have a better chance of finding a significant effect because it would better fit the likely distribution of the underlying data.

It is critical that researchers pay sufficient attention to analytic model comparison and selection since less appropriate models could greatly change a statistical inference and mislead researchers and policymakers into drawing potentially erroneous conclusions. In our fitted models and our sample of asthma patients, if the OLS results was used, it would be reasonable to conclude that the clinical information in an EHR has no impact on the number of laboratory tests ordered for the asthma patients who visited the ED. However, a more sophisticated and methodologically appropriate analysis demonstrates that it was associated with a difference. Therefore, comprehensive model checking is highly recommended to better understand the nature of observed data and to identify the most suitable statistical model.

Researchers are also encouraged to make comparisons across all the models discussed above whenever possible. Generally, PR is a starting point for count data analysis. When over-dispersion is present, NBR is considered instead. Since NBR is unable to account for over-dispersion that is due to excessive zeros, HR or ZIR models become the alternative modeling strategies. In our example, ZINBR is preferred to HR. However, different models may naturally fit different research questions. If we are interested in the number of allergy medications taken in the last spring semester by a random sample of college students, we would expect to have some students report taking zero allergy medications either because they never have allergies (in the CZG) or because they have allergies but not in last spring (in the NCZG). In this case, a ZIR model is more appropriate. In contrast, if we mainly focus on students who have ever taken allergy medications, a HR model would be preferred.

In practice, count data models are relatively easy to interpret and implement with the aid of statistical packages like SAS or Stata. However, we are aware that these techniques are not yet widely used by researchers in the biomedical informatics field. It is our hope that this discussion of count data modeling strategies

will lead to more frequent and appropriate use of such models in biomedical and health informatics research involving evaluations of HIT.

Acknowledgments We sincerely thank Dr Donald P Connelly from the University of Minnesota for his insightful comments. Dr Bryan Dowd was most helpful in initially suggesting that we investigate alternative statistical models for count data. We sincerely thank the staff from three partner health systems for their generous support in supplying the data.

Funding This project was funded in part under grant number UC1 HS16155 from the Agency for Healthcare Research and Quality, Department of Health and Human Services.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Cameron AC**, Trivedi PK. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press, 1998.
2. **Deb P**, Manning W, Norton E. *Modeling Health Care Costs and Counts*. 2006. <http://www.unc.edu/~enorton/DebManningNortonPresentation.pdf> (accessed 16 Apr 2011).
3. **Mosteller F**, Tukey JW. *Data Analysis and Regression: A Second Course in Statistics*. Reading: Addison-Wesley, 1977.
4. **Cohen J**, Cohen P, West SG, et al. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd edn. Mahwah, NJ: Lawrence Erlbaum, 2003.
5. **Vest JR**. Health information exchange and healthcare utilization. *J Med Syst* 2009;**33**:223–31.
6. **Rivera ML**, Donnelly J, Parry BA, et al. Prospective, randomized evaluation of a personal digital assistant-based research tool in the emergency department. *BMC Med Inform Decis Mak* 2008;**8**:3.
7. **Glassman PA**, Belperio P, Lanto A, et al. The utility of adding retrospective medication profiling to computerized provider order entry in an ambulatory care population. *J Am Med Inform Assoc* 2007;**14**:424–31.
8. **Turchin A**, Kolarik NS, Pendergrass ML, et al. Computational analysis of non-adherence and non-attendance using the text of narrative physician notes in the electronic medical record. *Med Inform Internet Med* 2007;**32**:93–102.
9. **Rosenbloom ST**, Geissbuhler AJ, Dupont WD, et al. Effect of CPOE user interface design on user-initiated access to educational and patient information during clinical care. *J Am Med Inform Assoc* 2005;**12**:458–73.
10. **Carroll AE**, Tarczy-Hornoch P, O'Reilly E, et al. The effect of point-of-care personal digital assistant use on resident documentation discrepancies. *Pediatrics* 2004;**113**:450–4.
11. **Scotch M**, Duggal M, Brandt C, et al. Use of statistical analysis in the biomedical informatics literature. *J Am Med Inform Assoc* 2010;**17**:3–5.
12. **Liu W**, Cela J. *Count Data Models in SAS*. 2008. <http://www2.sas.com/proceedings/forum2008/371-2008.pdf> (accessed 16 Apr 2011).
13. **Hausman J**, Hall BH, Griliches Z. Econometric models for count data with an application to the patents-R & D relationship. *Econometrica* 1984;**52**:909–38.
14. **Mullahy J**. Specification and testing of some modified count data models. *J Econom* 1986;**33**:341–65.
15. **Theera-Ampornpant N**, Speedie SM, Du J, et al. Impact of prior clinical information in an EHR on care outcomes of emergency patients. *AMIA Annu Symp Proc* 2009;**2009**:634–8.
16. **Kleinbaum DG**, Klein M. *Logistic Regression: A Self-Learning Text*. 3rd edn. New York, NY: Springer, 2010.
17. **Akaike H**. A new look at the statistical model identification. *IEEE Trans Automat Contr* 1974;**AC-19**:716–23.
18. **Schwarz G**. Estimating the dimension of a model. *Ann Stat* 1978;**6**:461–4.
19. **Long JS**, Freese J. *Regression Models for Categorical Dependent Variables using Stata*. 2nd edn. College Station, TX: Stata Press, 2006.
20. **Vuong QH**. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989;**57**:307–33.
21. **Greene WH**. *Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models*. 1994. http://www.stern.nyu.edu/ecc/wkpapers/POISSON-Excess_zeros-Selection.pdf (accessed 16 Apr 2011).
22. **Mekhjian H**, Saltz J, Rogers P, et al. Impact of CPOE order sets on lab orders. *AMIA Annu Symp Proc* 2003;**2003**:931.
23. **Hwang JI**, Park HA, Bakken S. Impact of a physician's order entry (POE) system on physicians' ordering patterns and patient length of stay. *Int J Med Inform* 2002;**65**:213–23.
24. **Mutimer D**, McCauley B, Nightingale P, et al. Computerised protocols for laboratory investigation and their effect on use of medical time and resources. *J Clin Pathol* 1992;**45**:572–4.
25. **Nightingale PG**, Peters M, Mutimer D, et al. Effects of a computerised protocol management system on ordering of clinical tests. *Qual Health Care* 1994;**3**:23–8.
26. **Neilson EG**, Johnson KB, Rosenbloom ST, et al. The impact of peer management on test-ordering behavior. *Ann Intern Med* 2004;**141**:196–204.
27. **Poley MJ**, Edelenbos KI, Mosseveld M, et al. Cost consequences of implementing an electronic decision support system for ordering laboratory tests in primary care: evidence from a controlled prospective study in the Netherlands. *Clin Chem* 2007;**53**:213–19.
28. **Smith BJ**, McNeely MD. The influence of an expert system for test ordering and interpretation on laboratory investigations. *Clin Chem* 1999;**45**:1168–75.
29. **Sanders DL**, Miller RA. The effects on clinician ordering patterns of a computerized decision support system for neuroradiology imaging studies. *Proc AMIA Symp* 2001;583–7.
30. **Tierney WM**, Miller ME, McDonald CJ. The effect on test ordering of informing physicians of the charges for outpatient diagnostic tests. *N Engl J Med* 1990;**322**:1499–504.
31. **Wang TJ**, Mort EA, Nordberg P, et al. A utilization management intervention to reduce unnecessary testing in the coronary care unit. *Arch Intern Med* 2002;**162**:1885–90.
32. **Wilson GA**, McDonald CJ, McCabe GP Jr. The effect of immediate access to a computerized medical record on physician test ordering: a controlled clinical trial in the emergency room. *Am J Public Health* 1982;**72**:698–702.
33. **Bates DW**, Kuperman GJ, Jha A, et al. Does the computerized display of charges affect inpatient ancillary test utilization? *Arch Intern Med* 1997;**157**:2501–8.