Published in final edited form as:

Med Care. 2017 March ; 55(3): e16–e24. doi:10.1097/MLR.000000000000156.

Use of systematic methods to improve disease identification in administrative data: The case of severe sepsis

Saeid Shahraz, MD PHD,

Heller School of Social Policy and Management, Brandeis University, 415 South Street, Waltham, MA 02453, USA, T: 6174174927 F: 7817363905

Tara Lagu, MD MPH,

Tufts University School of Medicine, 280 Chestnut St. 3rd FL, Springfield, MA 01199, USA, T: 4137947688, F: 4137948866, tara.lagu@baystatehealth.org; Center for Quality of Care Research, Baystate Medical Center, Springfield, MA; Division of General Medicine, Baystate Medical Center, Springfield, MA

Grant A. Ritter, PHD,

Heller School of Social Policy and Management, Brandeis University, 415 South Street, Waltham, MA 02453, USA, T: 7817363964 F: 7817363905, ritter@brandeis.edu

Xiadong Liu, Ed.D, and

Department of Psychology, Brandeis University, 415 South Street, Waltham, MA 02453, USA, T: 7817363244, F: 7817363905, xliu0806@brandeis.edu

Christopher Tompkins, PHD

Heller School of Social Policy and Management, Brandeis University, 415 South Street, Waltham, MA 02453, USA, T: 7817363964 F: 7817363905, Tompkins@brandeis.edu

Abstract

Background—Selection of ICD (International Classification of Diseases)-based coded information for complex conditions such as severe sepsis is a subjective process and the results are sensitive to the codes selected. We use an innovative data exploration method to guide ICD-based case selection for severe sepsis.

Method—Using the Nationwide Inpatient Sample, we applied Latent Class Analysis (LCA) to determine if medical coders follow any uniform and sensible coding for observations with severe sepsis. We examined whether ICD-9 codes specific to sepsis (038.xx for septicemia, a subset of 995.9 codes representing systemic inflammatory response syndrome, and 785.52 for septic shock) could all be members of the same latent class.

Results—Hospitalizations coded with sepsis-specific codes could be assigned to a latent class of their own. This class constituted 22.8% of all potential sepsis observations. The probability of an observation with any sepsis-specific codes being assigned to the residual class was near zero. The chance of an observation in the residual class having a sepsis-specific code as the principal

diagnosis was close to zero. Validity of sepsis class assignment is supported by empirical results, which indicated that in-hospital deaths in the sepsis-specific class were around 4 times as likely as that in the residual class.

Conclusion—The conventional methods of defining severe sepsis cases in observational data substantially misclassify sepsis cases. We suggest a methodology that helps reliable selection of ICD codes for conditions that require complex coding.

Keywords

LCA; observational data; severe sepsis; administrative discharge data

INTRODUCTION

Sepsis, a syndrome of infection complicated by host inflammatory response, is a frequent cause of morbidity and mortality in the US, with 700,000 hospitalizations and 200,000 deaths per year [1]. Severe sepsis is defined by the American College of Chest Physicians/ Society of Critical Care Medicine Consensus (ACCP/SCCM) as sepsis complicated by organ dysfunction. SCCM defines septic shock as severe sepsis with hypotension that requires vasopressors to maintain perfusion [2,3]. While the methods used to translate clinical data into diagnosis codes vary across medical coders and hospitals, the National Center for Health Statistics (NCHS) annually publishes the official guidelines for coding diseases and injuries, which details appropriate coding of sepsis syndromes [4].

Despite these coding standards, the best method for identifying cases of severe sepsis in the United States has been a subject of debate in the medical literature[3, 5-8]. Most published methods have applied various combinations of International Classification of Diseases (ICD) codes for infection and organ dysfunction to administrative datasets such as the Nationwide Inpatient Sample to calculate the number of cases annually. However, these methods have been criticized, because translating a pathophysiological description (infection with host inflammatory response and organ dysfunction) into administrative codes is an intricate and uncertain process [9]. Further complicating the issue is the fact that each method uses different code combinations, and the best method has not been agreed upon [3, 5-8].

These controversies have arisen because determining the sub-conditions which identify a medical syndrome is a subjective task, and different combinations of sub-conditions identify different cohorts of patients [3]. Figure 1 summarizes the sources of ambiguity in the process of translation from a clinical case of sepsis into an ICD-based diagnosis. First, it is unknown to what extent clinical definition and NCHS coding guidelines represent the same sepsis construct. Second, there is variation in how physicians apply the ACCP/SCCM definitions when recording the diagnosis of sepsis on medical charts. Third, there is a lack in uniformity in how coders apply the NCHS guidelines to read medical charts. Finally, researchers' choice in code selection for identifying sepsis is not consistent. These sources of uncertainty can lead to inconsistent identification of sepsis cases and thus in estimates of the burden of sepsis.

Given the considerable variability in the estimates of cost and prevalence of sepsis, there has been a call for a more uniform case definition [3, 5]. Historically, medical chart review has been used as gold standard to validate ICD-9-CM-based case identification [10-13], but extracting data from medical records is challenging, time-consuming and expensive [14]. In addition, even chart review cannot counter the problem of diagnostic uncertainty in identification of severe sepsis cases [2].

We propose a novel, quantitative approach existing sepsis definitions based on ICD codes with the goal of identifying the combination of ICD-9-CM codes that are most internally consistent. The method we will employ, Latent Class Analysis (LCA) has several assumptions. 1.) Sepsis is a latent construct, which is defined by a combination of ICD-9-CM codes; 2.) ICD-9-CM coding selection is subject to variation across coders, within and between hospitals, and therefore is an imperfect indicator of sepsis. We will use LCA to examine the consistency in coding for severe sepsis and to determine the set of highly correlated ICD-9-CM codes that can be used to identify the true cases of severe sepsis. This will open the door for a new estimation method that takes into account uncertainties and variation in the use of ICD codes in representing severe sepsis or any similar condition, allowing for uniform estimation of disease prevalence.

METHODS

Data

We combined the five most recent years (2007-2011) of Nationwide Inpatient Sample (NIS) datasets. NIS is the largest all-payer nationally representative hospital discharge data [15] and is commonly used to report annual prevalence of hospital based medical conditions in the United States. In order to work with a reasonably sized dataset, we used a 5% random sample of the combined data (n=1,991,830 observations). We then took a 5% random sample 5 times to ensure the consistency of the results.

Currently, two different ICD-9-CM code categories and a single code for septic shock are directly linked to diagnosis of sepsis. The first category consists of 17 subcodes of the 3-digit ICD code 038 (septicemia). A second category of codes for sepsis, the so-called Systemic Inflammatory Response syndrome (SIRS) codes, has been effective since October 1, 2002. SIRS code 995.91 represent sepsis without organ failure and 995.92 indicates sepsis in the presence of an organ failure [16]. Finally, beginning October 1, 2003 septic shock received a unique code, 785.52[17].

In 2001, Angus et al. suggested a method to match the pathophysiological description of the condition with ICD codes representing the condition [7]. Notably, this was prior to the introduction of the severe sepsis-specific and septic shock diagnosis codes. This first effort at identifying sepsis introduced a set of 109 ICD codes to represent either a bacterial or a fungal infection and another set of 13 codes to assign an organ failure. The condition of sepsis was identified as a code for infection and a second code for organ dysfunction [7]. Since this introduction, researchers and data analysts have frequently employed the “Angus” methodology to define cases of severe sepsis [18]. Some researchers have modified Angus’ method by addition/deletion of particular codes.

But the Angus definition is only one of several different methods of code selection (Table 1). As a result of differing selections of ICD codes, the variation in estimated prevalence of sepsis has ranged as much as 3.5-fold [5]. Notably, it appears that a larger set of infection codes is associated with a larger the estimate severe sepsis prevalence. In the published literature, two distinctly different prevalence estimates can be distinguished, based on the number of infection codes included for identification. Angus and Wang used a large set of infection codes, while Martin and Dombrovskiy limited included codes to those specific for sepsis rather than any infection. As a result, the estimates of Dombrovskiy et al.[8] and Martin et al.[19] were substantially lower than estimates made by Angus, Wang, and other researchers [7, 18].

To build the analytic file for this study, we identified all ICD diagnosis codes for infection and organ failure used by these four referenced authors (Angus, Wang, Martin, and Dombrovskiy) [7, 8, 18, 19]. Among the organ failure codes there were some procedure codes indicative of organ failure as well (see below). All inpatient stays with at least one infection diagnosis and one organ failure code from the combined the lists were included as observations in our analytic file. Since the NIS dataset allows for up to 15 diagnosis variables per stay, it is unlikely a sepsis code would not appear because of lack of sufficient room. The final ICD codes used for selection (infection, organ failure, and procedure codes) can be found from Appendix. The final analytic file consisted of 123,287 observations.

LCA as a method

LCA is a data exploratory technique that inputs observed categorical variables and predicts class membership on the basis of variables values. In LCA, class is an unobserved (latent) construct, which correlations among observed variables will reflect. LCA provides two key outputs (parameters): probability of class membership for each observation (conditional item probability parameter) and overall prevalence in each class (class probability parameter). Expectation Maximization Algorithm (EMA) is used to calculate the class membership likelihood. The process of likelihood estimation is iterative. In each iteration, the procedure algorithm attempts to improve the fit of the model such that expected class membership values get closer to observed class values based on the same identification codes. Choosing the appropriate number of distinct classes is a decision-making step in LCA. This decision is usually based on two criteria: statistical fit and substantive theory or interpretability of the classes. For LCA, methods such as Lo-Mendell-Rubin likelihood ratio test (LMR-LRT), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), and a specific measure of entropy are the best candidates to determine the appropriate number of classes. LMR-LRT value compares the improvement of fit across neighboring classes. Its associated P value shows if adding one more class causes the model to be a better fit to the data [21]. AIC or BIC with a lower value is indicative of better fit. Entropy is a measure ranges from 0 through 1. The higher the entropy measure the lower is the chance of case misclassification. The last but perhaps the salient important criterion is confirmation that the number of latent classes is sensible on the basis of substantive interpretability. This can be achieved by getting a desired distribution for each latent class, adequate number of individuals assigned to each class, and acceptable strength of association between classes and the observed characteristics (indicator variables). More details on LCA can be found elsewhere [22, 23].

Model building and analysis

The core to sensible criteria for identifying severe sepsis is to make appropriate rules for class membership using LCA. As a starting point, we used the NCHS guidelines, which require coders assign a code or set of ICD codes that together define the severe sepsis syndrome [4]. We aimed to test if selection of specific recommended sepsis codes (038.x subcodes, 785.52, and SIRS codes) by coders could cause an observation to be a member of a certain latent class versus any other less specified classes. In preparation for LCA, four binary (yes/no) indicators were created and added to the analytic file: 1.) an indicator variable that any subcode of ICD 9 CM code of 038.xx (indicator for code 038) was on discharge diagnosis list (DX1 through DX15); 2.) an indicator variable that any SIRS codes(indicator for SIRS) was placed on the diagnosis list; 3.) an indicator that the code 785.52 (indicator for septic shock) was on the diagnosis list and; 4.) an indicator for the presence of at least one of three procedures (indicator for procedure) associated with sepsis: electroencephalography, hemodialysis, and mechanical ventilation on the procedure list. We used a routine within MPLUS 7.1 to perform LCA and estimate resulting class membership fit. We performed several runs specifying different number of classes in an effort to identify the most appropriate model based on substantive theory coupled with acceptable fit statistics and entropy (class predictability). Acceptable class predictability was determined to be a result with entropy greater than 0.8 [22].

We created four new variables to re-identify severe sepsis cases from each of the four previously published definitions. We compared the case identification by each of the definitions with that of LCA-defined cases for each individual in the dataset. Hence, we reported falsely identified cases for each of the four previous methods of case identification assuming LCA method as criterion. In the analysis and reporting the results we closely observed the ethical conducts in use of NIS data.

RESULTS

Repeated 5% samples provided LCA results, which were different at the third decimal position. We report the results of one of the samples at random. A total of 123,287 observations were in the final dataset. Indicators for code 038, SIRS, septic shock, and procedure were positive in 22.3%, 19.5%, 7.6%, and 6.6% of the observations respectively. LCA results revealed that 2-class and 3-class attempts returned interpretable and consistent solutions (Table 2), but a 4-class attempt proved to be unidentifiable. Based on fit statistics, the 3-class model is better than the 2-class solution. It provides a lower, non-significant Pearson Chi-Square Fit Statistic and lower AIC and BIC values indicating higher likelihood associated with its LCA parameters. In the 2-class solution, the more prevalent class, named class 1, had 77.2% of observations, and the less prevalent class, class 2, had 22.8%. Within class 2, 98% of observations included a 038 subcode on their diagnosis list, 85% included a SIRS code, and 33% included a septic shock code. The probability of a class 2 observation having a code from the sepsis procedure code list was 13%. In contrast, all observations in class 1 lacked a sepsis specific code from any of the three diagnosis sets, and only 5% of them had a sepsis related procedure code. Turning to the calculated indicators for the diagnosis sets, the odds ratio of being classified in class 2 versus class 1 was statistically

significant for all four ($p<0.001$). In light of these results, class 2 is referred to as the 'specifically coded class' and class 1 as the 'residual class'. The 3-class solution appears to keep class 1 the same, but split the original class 2 from the 2-class solution into two classes with class 3 representing the more severe cases. In the 3-class solution, cases in class 3 have a higher probability of having one of the procedure codes (21% in class 3 versus only 8% in class 2) and a higher probability of being coded as septic shock (59% in class 3 versus only 19% in class 2), but a lower probability of having a SIRS code (74% in class 3 versus 100% in class 2).

The principal diagnosis variable hosted 3,280 and 1,364 different ICD 9 CM codes for the class 1 (residual) and class 2 (specifically coded) respectively (Figure 2). Pneumonia, chronic obstructive pulmonary disease (COPD), and respiratory failure comprised 27.5% of principal diagnosis. In the specifically coded class, the principal diagnosis was a septicemia subcode for more than half of observations.

Further evaluation of the 2-class solution showed that the odds of in-hospital death was 4.5 times (95% CI: 4.4-4.6) greater for observations in the specifically coded class than that of the residual class. In the same vein, average length of stay for was 14.4 (95% CI: 13.22-14.58) days for specifically-coded class and 9.1 (95% CI: 8.95-9.16) days for the residual class. Similar results, though less striking (odds ratio=2, with 95 CI between 1.8 and 2.2), were obtained for the probability of being coded by one of the ICD codes that explicitly refer to a systemic infection (e.g. code 112.5 for systemic candidiasis).

Table 3 shows the extent to which case identification can be misclassified if one adopts different case selection methods offered by previous authors and case identification by LCA as reference. Classification by Angus et al. and Wang et al. yielded a low specificity and positive likelihood ratio, and as a result a low fraction of correctly classified cases. Conversely, classification by Dombrovsky et al. and Martin et al. yielded a high specificity and positive likelihood ratio and thus fraction of cases correctly classified through these two methods was acceptably high.

DISCUSSION

Using a broad definition of severe sepsis that employed any infection code plus an organ dysfunction code, we identified hospitalizations meeting this criteria and used latent class analysis to separate cases into two classes: a class identified by specifically coded ICD codes [4] and a class that met the infection plus organ failure criteria but was almost void of ICD-9-CM codes specific for sepsis. Based on LCA's algorithm for maximizing estimation, most cases of LCA-defined severe sepsis are coded following Dombrovskiy or Martin recommendations for case identification [8]. Using a broader definition of severe sepsis markedly reduces the correlation between estimated and observed sepsis, and may overestimate the number of severe sepsis cases by as high as two- to four-fold, depending on how the LCA model was specified. This observation is also compatible with our results of severe sepsis prevalence in Table 3 in which the prevalence of LCA-defined cases is very close to that of Dombrovskiy and Martin. We also found that hospitalizations identified as sepsis using specific codes had a relatively homogenous set of principal diagnoses, while the

set of principle diagnoses for cases in the residual class was widely heterogeneous. The incidence of in-hospital death for the specifically coded class was almost 4 times higher than that for the residual class, demonstrating that the two classes had very different risk profiles. Finally, we illustrated that case misclassification was greatly lower in previous literature when a more specified method of case identification was used.

Medical diagnosis is a subjective matter. Some conditions such as cancers have specific diagnostic criteria that make them less controversial for recording and reporting purposes. Other conditions, including sepsis, have no gold standard definition against which diagnostic criteria can be calibrated [24]. To complicate the situation further, clinical definitions of sepsis lack sufficient certainty [2]. These components of error can result in case misclassification, [14] with implications for research, planning, and evaluation [14].

Using exploratory data techniques such as LCA to identify sepsis cases in observational databases is a novel but not unknown approach. Rindskopf et al. and Faraone et al. emphasized the value of LCA in medical case identification in the absence of a gold standard [25, 26]. LCA has also been successfully implemented for claims-type data in previous research. Rector et al. showed that using multiple criteria (algorithmic case identification) improves both sensitivity and specificity of case identification in a set of claim-based data [27]. Favorable results were obtained using LCA model to diagnose depression [28]. Prosser et al. found out that conventional case definition for asthma using administrative data would end up with an underestimation of cases and suggested LCA as an alternative case identification tool[14].

Still, the results of this study should be judiciously interpreted. First, our results depend on clinical diagnoses coded in claims data. If a physician has made the wrong diagnosis and document that diagnosis in the chart, it can be translated into a wrong code. Second, we are not able to detect upcoding, a type of administrative misconduct that occurs when an ICD code is intentionally replaced by an alternative ICD code that is financially more lucrative in terms of provider reimbursement [29, 1, 30]. Third, LCA-based case identification is not absolute and comes with some degrees of uncertainty around case determination [28], even if all ICD codes truly represent the cases. This means that even if sepsis-specific codes are a true proxy of what they intend to be, the class membership is still a non-definitive process. However, the class membership probability in our study came with minimal statistical uncertainty and the results were fairly stable.

We suggest two potential case selection methods for severe sepsis. The first is to rely strictly on the results of LCA. Class 2 of the 2-class solution can be used to define these cases. The second way is to decide on a definitive case identification method guided by LCA analysis. The definition might be as follows: any observation is a severe sepsis if it is coded with a combination of an organ failure code or its SIRS equivalent (995.92) and at least one of the two sets of codes: an ICD 9 CM code from the 038.xx code group (septicemia) or the specific code for septic shock (785.52). Either of these two selection methods would prove to unify and make more homogeneous the identification of severe sepsis. While we used the 2-class solution as a guide for code selection, we can also suggest that the 3-class results can be also used in two ways: 1) to divide cases into two severity classes if needed, and 2) as a

validity check for the model entire analysis such that the third class must be a sub-class of the second class with interpretable results.

Research is also needed to unravel the extent to which use of NCHS' official recommendations is sensitive and specific. This can be done by sepsis identification through extensive medical chart review and comparison with identification based on the method we proposed or an improvement of it. To undertake this task, however, clear and reproducible guidelines for sepsis diagnosis based on chart review are needed. NCHS official guidelines entirely rely on recorded clinical diagnosis in the chart, but none of the previous validation studies using chart review have declared how severe sepsis should be identified through the chart review process.

This study has important implications. Prior sepsis definitions have used intuitive definitions and then attempted to validate those definitions through a series of analyses using administrative data and chart review. In contrast, this analysis prompts researchers and data analysts to forgo methods that solely rely on expert opinion for developing these definitions and instead replace them with a systematic data-driven and replicable method such as LCA. The introduction of ICD-10 offers the prime opportunity for the application of a systemic methodology, because ICD 10 CM contains a much larger number of codes than the current version of ICD and uses up to 7 digits of codes (compare this to 5-digit maximum resolution of ICD 9 system). This is true not just for severe sepsis but for any medical condition with complex coding guidelines for case identification. Such a change could have a significant impact on estimation of the burden of disease and the epidemiology of many complex medical conditions.

Acknowledgments

Funding Disclosure

This work, as part of the PHD dissertation of the first author, has been financially supported jointly by Agency for Healthcare Research and Quality (AHRQ) and Schneider Family of Schneider Institute of Health Policy (SIHP). Dr. Lagu is supported by the National Heart, Lung, and Blood Institute of the National Institutes of Health under Award Number K01HL114745

REFERENCES

1. Lagu T, Rothberg MB, Shieh MS, Pekow PS, Steingrub JS, Lindenauer PK. Hospitalizations, costs, and outcomes of severe sepsis in the United States 2003 to 2007. *Crit Care Med.* 2012; 40:754–61. [PubMed: 21963582]
2. Bone RC, Balk RA, Cerra FB, Dellinger RP, Fein AM, Knaus WA, et al. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest.* 1992; 101:1644–55. [PubMed: 1303622]
3. Wilhelms SB, Huss FR, Granath G, Sjöberg F. Assessment of incidence of severe sepsis in Sweden using different ways of abstracting International Classification of Diseases codes: difficulties with methods and interpretation of results. *Crit Care Med.* 2010; 38:1442–9. [PubMed: 20400903]
4. Statistics, NCIH, editor. Centers for Diseases Control and Prevention; Georgia, Atlanta: 2006. ICD-9-CM Official Guidelines for Coding and Reporting.
5. Gaieski DF, Edwards JM, Kallan MJ, Carr BG. Benchmarking the incidence and mortality of severe sepsis in the United States. *Crit Care Med.* 2013; 41:1167–74. [PubMed: 23442987]

6. Lagu T, Rothberg MB, Shieh MS, Pekow PS, Steingrub JS, Lindenauer PK. What is the best method for estimating the burden of severe sepsis in the United States? *J Crit Care*. 2012; 27:414 e1–9. [PubMed: 22516143]
7. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit Care Med*. 2001; 29:1303–10. [PubMed: 11445675]
8. Dombrovskiy VY, Martin AA, Sunderram J, Paz HL. Rapid increase in hospitalization and mortality rates for severe sepsis in the United States: a trend analysis from 1993 to 2003. *Crit Care Med*. 2007; 35:1244–50. [PubMed: 17414736]
9. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005; 40(5 Pt 2):1620–39. [PubMed: 16178999]
10. Cooke CR, Joo MJ, Anderson SM, Lee TA, Udris EM, Johnson E, et al. The validity of using ICD-9 codes and pharmacy records to identify patients with chronic obstructive pulmonary disease. *BMC Health Serv Res*. 2011; 11:37. [PubMed: 21324188]
11. Hjerpe P, Merlo J, Ohlsson H, Bengtsson Bostrom K, Lindblad U. Validity of registration of ICD codes and prescriptions in a research database in Swedish primary care: a cross-sectional study in Skaraborg primary care database. *BMC Med Inform Decis*. 2010; 10:23.
12. Stein BD, Charbeneau JT, Lee TA, Schumock GT, Lindenauer PK, Bautista A, et al. Hospitalizations for acute exacerbations of chronic obstructive pulmonary disease: how you count matters. *COPD*. 2010; 7:164–71. [PubMed: 20486814]
13. Bozic KJ, Chiu VW, Takemoto SK, Greenbaum JN, Smith TM, Jerabek SA, et al. The validity of using administrative claims data in total joint arthroplasty outcomes research. *J Arthroplasty*. 2010; 25(6 Suppl):58–61. [PubMed: 20570479]
14. Prosser RJ, Carleton BC, Smith MA. Identifying persons with treated asthma using administrative data via latent class modelling. *Health Serv Res*. 2008; 43:733–54. [PubMed: 18370976]
15. Healthcare Cost and Utilization Project(HCUP). HCUP Databases. Agency for Healthcare Research and Quality; Rockville, MD: 2007-2011.
16. AHA Coding Clinic AHA Coding Clinic Chicago. Illinois American Heart Association Fourth Quarter; 2003. p. 80
17. AHA coding Clinic Chicago. Illinois American Hospital Association Fourth Quarter; 2003.
18. Wang HE, Shapiro NI, Angus DC, Yealy DM. National estimates of severe sepsis in United States emergency departments. *Crit Care Med*. 2007; 35:1928–36. [PubMed: 17581480]
19. Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. *N Engl J Med*. 2003; 348:1546–54. [PubMed: 12700374]
20. Collins, LML. S.T. . Latent class and latent transition analysis, with applications in social, behavioral, and health sciences. Wiley & Sons; Hoboken, New Jersey: 2010.
21. Nylund KA, T. Muthen B. Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Struct Equ Modeling*. 2007; 14:35. 14., 2007.
22. McCutcheon, AL. Latent class analysis. Sage Publications; Newbury Park: 1987. p. 96
23. Hagenaars, JA., McCutcheon, AL. Applied latent class analysis. Vol. xxii. Cambridge University Press; Cambridge; New York: 2002. p. 454
24. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Intensive Care Med*. 2003; 29:530–8. [PubMed: 12664219]
25. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. *Stat Med*. 1986; 5:21–7. [PubMed: 3961312]
26. Faraone SV, Tsuang MT. Measuring diagnostic accuracy in the absence of a “gold standard”. *Am J Psychiatry*. 1994; 151:650–7. [PubMed: 8166304]
27. Rector TS, Wickstrom SL, Shah M, Thomas Greenlee N, Rheault P, Rogowski J, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. *Health Serv Res*. 2004; 39(6 Pt 1):1839–57. [PubMed: 15533190]

28. Garrett ES, Eaton WW, Zeger S. Methods for evaluating the performance of diagnostic tests in the absence of a gold standard: a latent class model approach. *Stat Med.* 2002; 21:1289–307. [PubMed: 12111879]
29. Brunt CS. CPT fee differentials and visit upcoding under Medicare Part B. *Health Econ.* 2011; 20(7):831–41. [PubMed: 20681033]
30. Hsia DC, Krushat WM, Fagan AB, Tebbutt JA, Kusserow RP. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med.* 1988; 318:352–5. [PubMed: 3123929]

Figure 1.
The process of ICD code selection for severe sepsis

Figure 2.

This figure shows the percentage of principal diagnosis (DX1) for observations in Class 1 (bottom chart with blue bars) and that of Class 2 (top chart with gray bars) in the 2-Class solution LCA.

Table 1

Method of ICD code selection to define sepsis by different researchers Please note that different authors use the ICD codes at some different granularity level (3-digit vs. 4- or 5-digit), which may partly contribute to the number of codes selected.

	038 (septicemia) sub-codes and Septic shock (785.5)	No. Infection codes	No. Organ Failure codes	SIRS 995.90	SIRS 995.91	SIRS 995.92
Angus et al.	YES	109	13	NO	NO	NO
Wang et al.	YES	108	14	YES	NO	YES
Martin et al.	YES	5	29	NO	NO	NO
Dombrovskiy et al.	YES	8	22	NO	YES	YES

Table 2

Abbreviations used: ICD: International Classification of Disease, SIRS: Slow Inflammatory Response Syndrome, AIC: Akaike Information Criteria, BIC: Bayesian Information Criteria, DF: Degrees of freedom, LMR LRT: Lo-Mendell-Rubin Likelihood Ratio Test Note: 123,287 cases of severe sepsis (at least one infection code and one organ failure code for each observation) were analyzed. Class prevalence is shown in percentage. Conditional probability for each indicator being classified in one of the available classes is indicated by proportions using two decimals. Observations can only be classified in one of the classes. Probabilities less than 0.01 were rounded down to zero and those above 0.99 rounded up to 1.

Indicators:	2-class solution (class prevalence %)		3-class solution (class prevalence %)		
	Class 1 (77.2%)	Class 2 (22.8%)	Class 1(78.4%)	Class 2(12.8%)	Class 3 (8.8%)
<i>ICD code 038.xx</i>	0	0.98	0.15	0.97	0.98
<i>SIRS codes</i>	0	0.85	0.03	1	0.74
<i>ICD 785.52 (septic shock)</i>	0	0.33	0	0.19	0.59
<i>The procedure codes</i>	0.05	0.13	0.05	0.08	0.21
Entropy and Fit statistics:					
<i>Entropy (class predictability)</i>	0.99		0.89		
<i>Pearson Chi Squared</i>	314.5(DF=6, <i>P</i> value <0.001)		2.9(DF=1, <i>P</i> value <0.09)		
<i>AIC</i>	258771		258490		
<i>BIC</i>	258830		258635		
<i>LMR LRT</i>	118,246, <i>P</i> value=0.3		27,747, <i>P</i> value=0.001		

Table 3

This table shows prevalence of cases identified in the data set, sensitivity, specificity, misclassification percentage, and positive/negative likelihood ratio for 4 previous methods(Angus et al, Wang et al, Martin et al, and Dombrovskiy et al). Reference group is the results of the 2-class LCA analysis.

	Prevalence and 95% Confidence interval (%)	Sensitivity (%)	Specificity (%)	Correctly classified (%)	Positive Likelihood ratio	Negative Likelihood Ratio
Dombrovsky et al.	20.6(20.4-20.8)	90.1	99.7	97.5	283.7	0.1
Martin et al.	22.4(22.2-22.6)	87.4	96.5	94.4	24.8	0.13
Wang et al.	66.7(66.5-66.9)	84.7	38.5	48.9	1.4	0.4
Angus et al.	67.7(67.4-67.9)	84.6	37.2	47.9	1.3	0.4