A Comprehensive Assessment of Claims-Based Algorithms for Identifying Individuals with Diabetes

Changchun Wang MS[1], David Gilbertson PhD[1], Jiannong Liu PhD[1], Cheryl Arko BA[1], Shu-cheng Chen MS[1], Marshall McBean MD MSc[1,2], Allan Collins MD FACP[1,2]
[1]United States Renal Data System, Minneapolis Medical Research Foundation, [2]University of Minnesota Twin Cities

Introduction

- Hebert, et al. assessed a number of claims-based algorithms for identifying individuals with diabetes using Medicare Parts A and B claims (Am Jour of Medical Quality 14:6 270-277, 1999).
- Using the latest Medicare Current Beneficiary Survey (MCBS) 2001-2002 Access to Care data, we analyzed claims-based algorithms based on virtually all possible combinations of of diabetic claims in Inpatient hospital (IP), Skilled nursing facility (SNF), Home health (HH), Outpatient (OP) and Physician/supplier (PB) claim files.
- Using our weighting system, each algorithm was assessed by different combinations of "accuracy" characteristics (sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV)) (Table 1).
- Our objective was to find the most robust algorithms among the 299 proposed algorithms used to define claim-based diabetes.

Methods

- In order to ensure that two years of claims data were available to search for evidence of diabetes, selection criteria required that patients were:
  - at least 65 years of age at the start of the time period and resided in the 50 States or the District of Columbia,
  - in both the 2001 and 2002 MCBS surveys and alive before December 31, 2002, and,
  - were eligible for Medicare Parts A and B and never enrolled in an HMO during the two-year time period.
- The MCBS Access to Care file contains self-reported information on whether or not individuals have diabetes -this information was used as our "gold standard."
- 299 claims-based algorithms were proposed. Only "OR" relationships were considered among claim sources.
  - Sector I: 3^5-1 = 242 (5 sources: IP, SNF, HH, OP and PB, 3 categories for each source: 0, ≥1, ≥ 2)
  - Sector II: $2*3^3$ = 54 (5 sources, for combined OP and PB: 2 categories: ≥ 1, ≥ 2. For others: 3 categories: 0, ≥ 1, ≥ 2)
  - Sector III: 3 (pool all 5 sources together)
- Weighting system: under various considerations (Tables 2.1-2.4), 39 sets for sensitivity, specificity, PPV and NPV were proposed in the three sectors.
- Analytic methods
  - Two distances of characteristics of each algorithm to "perfect" algorithm (sensitivity = 1, specificity = 1, PPV = 1, and NPV = 1) were measured (Table 3).
  - 78 measurements (2 * 39) were calculated under 39 sets of weights and were sorted ascending distance.

Results

- Five common claims-based algorithms were singled out based on the top 58 algorithms in each measurement (Table 4).
- This study shows that the original methods proposed by Hebert et al. are nearly as accurate as the best algorithms found by our methodology.

Table 1
Statistics of each algorithm

		Self-reported Diabetes		
		Yes	No	
Claims-based Diabetes	Yes	TP (True positive)	FP (False positive)	PPV = TP/(TP+FP)
	No	FN (False negative)	TN (True negative)	NPV = TN/(TN+FN)
		Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

Table 2.1
Annotations in weighting system

N1: weighted number of samples with self-reported comorbidity
N2: weighted number of samples without self-reported comorbidity
N3: weighted number of samples with claim-defined comorbidity
N4: weighted number of samples without claim-defined comorbidity
W1: weight assigned to sensitivity
W2: weight assigned to specificity
W3: weight assigned to positive predicted value (PVD)
W4: weight assigned to negative predicted value (NVD)

Table 2.2
Weighting System: Part I
emphasizes sensitivity and specificity. Sum of W1 and W2 equals 100

W1	W2	W3	W4
25	75	0	0
30	70	0	0
35	65	0	0
40	60	0	0
45	55	0	0
50	50	0	0

Table 2.3
Weighting System: Part II
This sector considers all statistics together. PVD and NVD are treated equally important as sensitivity and specificity

W1	W2	W3	W4
5	45	5	45
5	45	10	40
5	45	15	35
5	45	20	30
5	45	25	25
10	40	5	45
10	40	10	40
10	40	15	35
10	40	20	30
10	40	25	25
15	35	5	45
15	35	10	40
15	35	15	35

W1	W2	W3	W4
15	35	20	30
15	35	25	25
20	30	5	45
20	30	10	40
20	30	15	35
20	30	20	30
20	30	25	25
25	25	5	45
25	25	10	40
25	25	15	35
25	25	20	30
25	25	25	25

Table 2.4
Weighting System: Part III
This sector considers all statistics together and use weighted number of samples as weights

W1	W2	W3	W4
N1	N2	0	0
sqrt(N1)	sqrt(N2)	0	0
N1+(N2-N1)/4	N2+(N1-N2)/4	0	0
N1+(N2-N1)/6	N2+(N1-N2)/6	0	0
Log(N1)	log(N2)	0	0
N1	N2	N3	N4
sqrt(N1)	sqrt(N2)	sqrt(N3)	sqrt(N4)
log(N1)	log(N2)	log(N3)	log(N4)

*sqrt( ): square root
log(): the natural (base e) logarithm

Table 3.
Distance to "perfect" algorithm

Distance I: Weighted geometric distance:
$\text{sqrt}\left(\sum Wi * (100 - Pi)^2\right)$

*∑: sum
Sqrt: square root
Wi: weight for ith statistic. i = 1, 2, 3, 4 for sensitivity, specificity, PVD and NVD.
Pi: value of ith statistic. i = 1, 2, 3, 4 for sensitivity, specificity, PPV and NPV

Distance II: Weighted absolute distance
$\sum \text{abs}\,(Wi * (100 - Pi))$
** abs: absolute value

Table 4. Top 5 algorithms to identify diabetics using MCBS 2001-2002 data

Algorithm	Sensitivity	Specificity	PPV	NPV
SNF>=1 or HH>=1 or OP>=2 or PB>=2	89.7	95.4	82.6	97.4
SNF>=2 or HH>=1 or OP>=2 or PB>=2	89.7	95.4	82.6	97.4
IP>=2 or HH>=1 or OP>=2 or PB>=2	89.7	95.3	82.5	97.4
IP>=2 or SN>=1 or HH>=1 or OP>=2 or PB>=2	89.8	95.3	82.5	97.4
IP>=2 or SN>=2 or HH>=1 or OP>=2 or PB>=2	89.8	95.3	82.5	97.4
*IP>=1 or SN>=1 or HH>=1 or OP>=2 or PB>=2	90.4	95.1	81.8	97.6

* proposed by Hebert et al.

Conclusions

- Based on statistic characteristics, selected claims-based algorithms could reflect patients' real status of diabetes.
- The methodology used is a systematic method that could be used to identify the best algorithms to validate any claims-based comorbid conditions where a "gold standard" is available for comparison.

Limitations

- Only "OR" situations are considered among 5 claim sources. Some "AND" relations should be explored in the future study.
- We assume self-reported diabetes status as the "gold standard".

USRDS Coordinating Center
www.usrds.org • NIH/NIDDK/DKUHD