ESTIMATION OF DISEASE-SPECIFIC COSTS IN HEALTH INSURANCE CLAIMS:

A COMPARISON OF THREE METHODS

Etsuji OKAMOTO* and Eiichi HATA

Objective To compare the accuracy and validity of three different methods (Proportional Disease Magnitude method [PDM] with two different magnitude estimations: arithmetic means with correction by the authors; Proportional Allotment Estimator [PAE] by Tango; Maximum Likelihood Estimator [MLE] also by Tango) for estimating disease-specific costs in health insurance claims.

Methods Application of the three methods to a computer-generated simulation dataset whose disease-specific costs were known and to actual outpatient claims whose disease-specific costs were unknown.

Outcome measures For simulation data, the accuracy was assessed by correlation between known disease-specific costs and estimated disease-specific costs by the three methods. For actual claims, concurrent validity was assessed by inter-method correlations between pairs of the two methods.

Results All three methods showed good agreement and accuracy with the simulation data but marked disagreement when they applied to actual claims. MLE yielded an aggregate total of disease-specific costs exceeding the actual total by 21.3% and showed negative disease-specific costs in 18 out of 154 categories. Inter-method correlations showed that PDM with PAE and MLE correlated most strongly ($R^2 = 0.9022$) while the least correlation was observed for PDM with arithmetic means and MLE ($R^2 = 0.6861$).

Conclusion MLE is not usable for claims analysis but PDM yielded good estimates with two different methods of magnitude estimation using actual claims.

Key words : proportional disease magnitude method, proportional allotment estimator, maximum likelihood estimator, health insurance claims, econometrics, simulation

I. Introduction

In 1996 Okamoto proposed a method to objectively estimate disease-specific costs in health insurance claims (hereafter, claims) with multiple diagnoses and christened it the "Proportional Disease Magnitude method" (hereafter, PDM)[1]. In 2003, it could be demonstrated through simulation that using disease-specific arithmetic means of per diem per-disease cost with appropriate correction, PDM achieved good validity[2].

Tango later proposed a new, but similar method, namely the Maximum Likelihood Estimator (hereafter, MLE) as well as the Proportional Allotment Estimator (hereafter, PAE)[3]. He also validated the accuracy of MLE and PAE using simulation data[4]. In the present study, we attempted to validate the accuracy of three methods using not only simulation but also actual claims data.

Estimation vs. distribution

Health insurance claims charge certain amounts for treating one or more diseases. If a claim contains a diagnosis of X, Y and Z, the amount can be estimated by multiple regression analysis (MRA) as the sum of regression coefficients (B values) for the costs of the respective diagnoses. The estimated values should be close to the "actual values" if not exactly equal. MRA is a method to estimate dependent variables by assigning regression coefficients to explanatory variables.

Now, suppose a claim with disease X, Y and Z has the cost of 10,000 yen, what amount was spent for treating disease Y? This time, explanatory variables and dependent variables are reversed. Since the

* National Institute of Public Health
Etsuji Okamoto, National Institute of Public Health, Department of Management Sciences, 2–3–6, Minami, Wako-shi, Saitama 351–0197, Japan
E-mail: atoz@niph.go.jp

cost of 10,000 is fixed, we can only distribute the 10,000 yen to the three different parts. For this purpose, an estimation method such as MRA is not appropriate. PDM is a distribution method to distribute dependent variables in proportion to magnitudes assigned to explanatory variables[5].

A crucial difference between estimation and distribution is that estimation can be validated while distribution cannot. Thus, estimation can be validated by comparing the estimated values and actual values: a method to estimate 11,000 yen for a claim of 10,000 yen is superior to another method to estimate 12,000 yen for the same claim because the difference is smaller. However, distribution is inherently arbitrary and there is no "right" distribution let alone a validating process. Someone might distribute 10,000 yen to three diagnoses equally, others might make division in proportion to regression coefficients obtained by MRA (this is possible only when all coefficients are positive). Any distribution is correct as long as the sum equals 10,000 yen. Distribution can only be validated with artificially generated simulation data in which the costs of the three diagnoses are known but such data differ from actual claims data.

When one wants to ascertain the disease-specific costs in claims, it becomes a matter of how one should distribute the cost of a claim into disease-specific costs, not estimating the cost of a claim from the diagnoses it contains. Regression coefficients derived from actual claims almost always yield negative values and hence cannot be used for distribution. The reason behind the numerous negative values is that the cost of a claim with multiple diagnoses is not simply a sum of disease-specific costs (unlike simulation data). A claim with 10 diagnoses may have a small cost while a claim with only one diagnosis may have an exorbitant cost. When faced with such irregularities, MRA minimizes the difference between estimated values and actual values of individual claims by assigning positive and negative coefficients to make both ends meet.

In contrast, simulation data are artificially generated by summing up disease-specific costs to yield the cost of a claim. Since the cost of a claim always equals the sum of disease-specific costs, the latter estimated by MRA for individual claims match well with the disease-specific costs distributed by PDM using regression coefficients for magnitude.

Criteria for validation

Disease-specific costs can only be predicted by the distribution method and not by estimation. Also, the distribution method cannot be validated with actual claims. Then how can one tell which distribution method is valid? The authors propose the following criteria:

- Necessary conditions—validity in simulation data

Methods must demonstrate high validity with artificially generated simulation data whose right disease-specific costs are known. The validity is evaluated in terms of the correlation between right answers and the results of the method. Ideally the regression line should be $y=x$ and $R^2=1$. Because simulation data have their right answers, results of any methods should converge on these.

- Satisfactory conditions—concurrent validity in actual claims

Ultimate validity must be demonstrated with actual claims data. However, actual claims data have no right answers. Under this situation, concurrent validity must be a practical solution: applying different methods to the same data and see if the two results converge. If they do, it is plausible that the two methods are both valid. Unfortunately, there is no way to tell which one or both are invalid if they do not.

Characteristics of the three methods

Of three methods compared in this article, PDM with arithmetic means and PAE are both distribution methods. They are also similar in that both assume common values (magnitudes) for each diagnostic category. PDM calculates magnitudes in a rapid manner: calculating arithmetic means and correcting them with a formula. PAE calculates magnitudes in a step-by-step manner: repeating calculation of arithmetic means until the values converge.

MLE is an estimation method similar in some respects to MRA, differing in that it attempts to estimate disease-specific costs in individual claims by repeating the procedure until the values converge. As with MRA, MLE inevitably produces numerous negative disease-specific costs. It would be hard for any health professional to comprehend this concept. More critically, the disease-specific costs estimated by MLE will not sum up to the actual total cost of entire claims.

The relationships among the four methods, including MRA, are summarized in Figure 1.

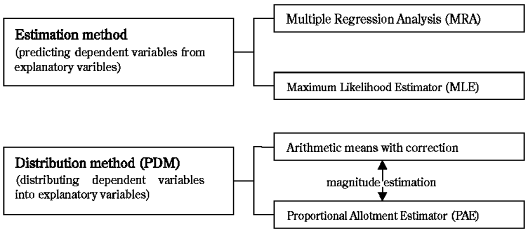
Estimation method
(predicting dependent variables from explanatory varibles)
Multiple Regression Analysis (MRA)
Maximum Likelihood Estimator (MLE)
Distribution method (PDM)
(distributing dependent variables into explanatory variables)
Arithmetic means with correction
magnitude estimation
Proportional Allotment Estimator (PAE)

Figure 1. Relationships among the different methods

II. Theory

A health insurance claim contains the following data.

- Cost (expressed in a monetary sum)
- The number of days (inpatient days for inpatient claims and the number of office visits for outpatient claims)
- Diagnoses (one or more). (They are coded in numerically corresponding to each diagnostic category. For example, hypertension is coded 901. For classification of health insurance claims in Japan, the so-called 119 classification system is typically used[6].)

The cost and the number of days are resources spent for treating the diagnosed diseases. However, no correspondence is given in claims on how much of the cost and the number of days were spent for each diagnosis. The three methods are intended to estimate those unobservable disease-specific costs and days in entire claims.

Let individual claims be denoted by *i* and diagnostic categories by *j*. Also, the total number of claims, the total costs, the total number of days and the total number of diagnoses in the entire claims are denoted R, P, D and N, respectively.

- Observable data

Pi, Di and Ni denote the cost, the number of days and the number of diagnoses in the *i*th claim ($1 \leqq i \leqq R$). Nj and Nij denote the number of diagnoses of the *j*th category (given 119 categories, $1 \leqq j \leqq 119$) in the entire set of claims and the number of diagnoses in the jth category in the ith claim. These are the observable data from which we have to estimate the following unobservable data.

- Unobservable data

Pj and Dj are the disease-specific cost and the number of days in the entire set of claims attributable to the *j*th category: the subjects of the estimation. Likewise, Pij and Dij denote the cost and the number of days in the *j*th category in the *i*th claim.

Their relationship is summarized as follows

$$P = \sum_{i=1}^{R} Pi = \sum_{j=1}^{119} Pj = \sum_{j=1}^{119} \sum_{i=1}^{R} Pij \qquad (1)$$

$$D = \sum_{i=1}^{R} Di = \sum_{j=1}^{119} Dj = \sum_{j=1}^{119} \sum_{i=1}^{R} Dij \qquad (2)$$

$$N = \sum_{i=1}^{R} Ni = \sum_{j=1}^{119} Nj = \sum_{j=1}^{119} \sum_{i=1}^{R} Nij \qquad (3)$$

The three methods estimate unobservable Pij and Dij from observable Pi, Di, Nij and Nj. If Pij and Dij are estimated, we will be able to obtain disease-specific cost Pj and days Dj simply by summing them up:

$$Dj = \sum_{i=1}^{R} Dij \quad Pj = \sum_{i=1}^{R} Pij \qquad (4)$$

PDM (Okamoto)

PDM is a distribution method that assumes a common magnitude for cost and days in each diagnostic category, expressed as $\dot{P}j$ and $\dot{D}j$ for the *j*th category, and that relative relationship among magnitudes of different categories in a claim are assumed constant, then Pij and Dij can be estimated as follows:

$$Dij = \frac{Nij * \dot{D}j}{\sum_{j=1}^{119} (Nij * \dot{D}j)} * \dot{D}i$$

$$Pij = \frac{Nij * \dot{P}j}{\sum_{j=1}^{119} (Nij * \dot{P}j)} * \dot{P}i \qquad (5)$$

1) Estimation of magnitude ($\dot{P}j$)

Okamoto first used the likelihood of becoming a primary diagnosis in each category obtained from Patient Survey as magnitude, and the authors demonstrated that arithmetic means of per diem per disease cost (hereafter, P/DN) with correction yielded good estimates. Tango proposed an iterative method called PAE and the authors regard it as yet another method of magnitude estimation for PDM. The following is an explanation of the two methods for magnitude estimation. Here we focus only on estimation of cost and do not deal with days.

(1) Arithmetic means with correction (Okamoto & Hata)

For magnitude estimation, we used arithmetic means of P/DN. Dividing by the number of days and diagnoses, we can minimize the cost-inflationary effect of days and diagnoses. For example, a claim of 2000 yen with 5 days and 4 diagnoses, P/DN = 100. We first calculate overall average of P/DN and disease-specific P/DNj as follows:

$$P/DN = \frac{P}{\sum_{i=1}^{R} (Di * Ni)} \qquad (6)$$

$$P/DNj = \frac{\sum_{i=1}^{R} \frac{Pi * Nij}{Di * Ni}}{\sum_{i=1}^{R} Nij} \qquad (7)$$

However, the arithmetic mean of a disease is diluted by other diagnoses. If diagnosis *j* has a magnitude higher than the overall average by $\varDelta P (= P/DN + \varDelta P)$, the observed average of a claim with disease j(P/DNj) would be $P/DN + \varDelta P/\mathbf{n}$ if the number of diagnoses in the claim is **n**. To estimate Pj from the overall average P/DN and the observed P/DNj, the following correction proved effective:

$$\dot{P}j = P/DNj * \left(\frac{P/DNj}{P/DN}\right)^{c} \qquad (8)$$

We demonstrated through simulation that the best validity is achieved when c(correction) = 2. We later discovered that c can be generalized as **c = ln (n)** when P/DNj > P/DN and **c = n − 1** when P/DNj < P/DN, where n is the average number of diagnoses in claims with *j*th diagnosis, which will be published elsewhere[7].

(2) Iterative method (Tango's PAE)

Set the initial value of $\dot{P}j$ as

$$Pij^{(0)}=\frac{Pi}{Ni} \text{ and } \dot{P}j^{(0)}=\frac{\sum_{i=1}^{R}(Nij*Pij^{(0)})}{Nj} \tag{9}$$

And repeat the process until $\dot{P}j^{(k)}$ converges to yield $\dot{P}j$.

$$Pij^{(k)}=\frac{Pi*\dot{P}j^{(k-1)}}{\sum_{j=1}^{119}(Nij*\dot{P}j^{(k-1)})} \tag{10}$$

and

$$\dot{P}j^{(k)}=\frac{\sum_{i=1}^{R}(Nij*Pij^{(k)})}{Nj} \tag{11}$$

$\dot{P}j$ so obtained will be applied to formula (4) and (5) to yield Pj.

MLE (Tango)

MLE directly estimates Pij through iteration and differs from PDM in that it does not distribute within an individual claim. It is also an iterative method but assumes a common variance in cost (σ^2).

Initial values for Pij and $\dot{P}j$ are the same as those in formula (9) and repeat the following procedures until $Pij^{(k)}$ converges. Pij so obtained will yield Pj using formula (4).

$$Pij^{(k)}=\dot{P}j^{(k-1)}+\lambda i*\sigma^2 \tag{12}$$

Where

$$\lambda i=\frac{\frac{Pi}{Di}-\sum_{j=1}^{119}(Nij*\dot{P}j^{(k-1)})}{\sigma^2*\sum_{j=1}^{119}Nij} \tag{13}$$

And

$$\dot{P}j^{(k)}=\frac{\sum_{i=1}^{R}(Nij*Pij^{(k)})}{\sum_{i=1}^{R}Nij} \tag{14}$$

III. Methods

PDM using magnitude of arithmetic means with correction was conducted with a computer program "PDM Ver. 2" which was produced by the authors with a research grant and was placed on the web as freeware (http://resept. com). Iterative procedures to estimate magnitudes by PAE and MLE were conducted with C++ language and a regular Windows PC.

Validation with simulation data

To validate the accuracy of the three methods, we applied them to simulation data whose disease-specific costs are known.

The simulation data consist of 1000 computer-generated health insurance claims, which mimic actual outpatient claims of Japan in terms of the number of days, case-mix of diagnoses and cost. The distribution of the number of diagnoses in a claim was set according to a published survey[8] with a maximum of 15 diagnoses in a claim.

Each diagnosis recorded in a claim was assigned a P/DN cost randomly generated in normal distribution with a mean reflecting disease-specific per diem cost obtained from a published survey[9] and a standard deviation of 30% of its value (i.e. coefficient of variance is set at 30%). Then P/DN costs assigned to all diagnoses in a claim are summed up to yield the cost of the claim.

The number of days is also assigned to each diagnosis in the same manner but the numbers of days assigned to all diagnoses in a claim are not summed up, instead the largest number of days out of them is chosen as the number of days of the claim. This reflects the assumption that the cost spent to treat each diagnosis will add up but the number of days (=number of office visits) will not simply add up, instead it will be equal to that for the diagnosis requiring the most frequent office visits. For example, a patient with diseases A, B and C, which require 4, 3 and 2 times a month, respectively, will only need to visit a doctor 4 times instead of 9.

Because of this assumption, the simulation data cannot be used for validation of estimation of the number of days because actual disease-specific days in a dataset of claims can not be known.

Specifications of the simulation data are as follows.

Number of claims: 1000
Total number of days: 2750 days
Total cost: 8,334,411 yen
Total number of diagnoses: 3,870
Total (day*number of diagnoses): 12,288
Average number of days per claim: 2.75 days
Average P/DN cost: 678.3 yen
Diagnostic categories: The standard 119 classification system was used, but there were no diagnoses in 10 out of 119 categories leaving the number of categories at 109.

Concurrent validity using actual claims

Simulation data are fictitious data artificially generated under certain assumptions. Therefore, validation with simulation data will not automatically translate into applicability to practical settings. However, there is no way to validate the accuracy when the methods are applied to actual claims because actual disease-specific costs can never be known.

Nevertheless, we here applied the three methods to actual claims and examined their concurrent validity[10]. Although there is no way to objectively measure which one of these methods is better than the other, the inter-method correlations can provide some insights as to how accurate they are.

The data used were outpatient claims submitted to Natori city's National Health Insurance program in February 2002. The data were intended for evaluation of influenza vaccination program by the city government, which was approved by the Ethics Review Committee of NIPH (NIPH-IBRA#03002) for analysis by PDM and its results were already published[11]. Data were provided to the authors in an unlinkable anonymous fashion pursuant to the city's Influenza Vaccination Appraisal Ordinance. This analysis was performed as part of an approved epidemiological study to ascertain the accuracy of estimation for disease-specific costs of influenza.

Specifications of the data were as follows:

Number of claims: 15,771

Total number of days: 32,695 days

Total cost: 209,754,920 yen

Total number of diagnoses: 59,330

Total (day*number of diagnoses): 138,096

Average number of days per claim: 2.07 days

Average P/DN cost: 1519.9 yen

Diagnostic category: In addition to the standard 119 classification system, Natori city added 41 mutually exclusive categories making the total number of categories to 160. There were no diagnoses in six categories leaving the number of categories at 154.

IV. Results

Validation using simulation data

PAE reached convergence at the 273rd iteration and MLE at the 237th iteration. The results are presented in Table 1. The aggregate total of disease-specific costs estimated by MLE was smaller than the actual total by 0.8%. The correlations of the results of three methods against "right answers" of the simulation data were as follows. All three methods faired well in terms of accuracy with simulation data.

PDM with arithmetic means

$y = 0.9837x + 1246.0 \quad R^2 = 0.9926$

PDM with PAE

$y = 0.9862x + 1057.3 \quad R^2 = 0.9956$

MLE

$y = 0.9659x + 2013.7 \quad R^2 = 0.9935$

Concurrent validity using actual claims

PAE reached convergence at the 618th iteration and MLE at the 318th iteration. The results with the three methods are presented in Table 2. MLE yielded negative values in 18 out of 154 categories and its aggregate total of disease-specific costs exceeded the actual total by 21.3% (254,448,765 yen vs. actual 209,754,920 yen). PDM with magnitudes estimated by PAE (PDM with PAE) yielded zero values in 10 out of 154 categories. Scatter grams showing mutual correlation among three methods are shown in Fig 2-4. PDM with PAE and MLE were found to be most strongly correlated ($R^2 = 0.9022$) while PDM with arithmetic means and MLE were correlated least ($R^2 = 0.6861$).

V. Discussion

All three methods demonstrated a good agreement in estimating the "right answers" in simulation data and one can safely conclude that all fulfilled the necessary conditions for validity. However, when they were applied to actual claims they showed disagreement. As discussed in section 1.1, costs of actual claims are not simply the sums of disease-specific costs, reflecting the irregularity of claims data.

It is something like solving exam questions. For questions with right answers, any good students will reach the same answers. However, for irregular questions with no definite right answers, no two students agree with their answers. At best, one can assume that right answers may be around where many students agree most.

Out of three methods, the authors consider MLE to be unsuitable for claims analysis for the following reasons: 1) it yielded negative disease-specific costs for numerous diagnostic categories and 2) the aggregate total of disease-specific costs estimated by MLE did not match the actual total (the estimate exceeded the actual one by 21.3% in claims and there was a slight underestimation with simulation data).

In comparison to MLE, PDM was able to estimate disease-specific costs with both methods of magnitude estimation (arithmetic means with correction and PAE). Concurrent validity was demonstrated by both methods ($y = 0.8337x$ and $R^2 = 0.8353$), suggesting that right answers should lie somewhere around the two results. It is safe to conclude that PDM, with whichever magnitude, fulfilled the conditions for satisfactory validity.

Still, results of PDM with magnitudes by PAE yielded zero disease-specific costs in numerous diagnostic categories: a questionable phenomenon given the nature and purpose of diagnoses in claims. Claims are financial documents and not medical certificates: diagnoses written in claims are intended to justify the treatment cost and not to merely certify that the patient has the disease. Therefore, the

Table 1. Validation using Simulation Data

Serial number	Diagnostic categories	right answer	PDM		MLE
			with arithmetic means	with PAE	
1	Intestinal infectious diseases	70938	69722	68474	66809
2	Tuberculosis	16771	10803	13487	15604
3	Sexually Transmitted Diseases	15541	21835	29458	27599
4	Viral infections with skin lesions	52420	56090	59489	59793
5	Viral hepatitis	74527	68851	73283	72315
6	Other viral disease	9637	18090	12320	10723
7	Mycoses	48287	59881	59242	56305
9	Other infectious diseases	6163	5975	6249	6927
10	Stomach cancer	90029	99047	87544	89597
11	Colon cancer	53681	43326	55260	56934
12	Rectal cancer	28796	28806	26069	25389
13	Liver cancer	20849	9057	8194	10310
14	Lung cancer	59411	64135	65299	54624
15	Breast cancer	65234	60565	58081	61099
16	Uterine cancer	11129	10769	12722	12101
17	Malignant lymphoma	14536	7416	11308	12727
18	Leukemia	9946	13035	13412	13092
19	Other malignant neoplasms	111968	117950	109291	119561
20	Benign neoplasm	209632	200654	202549	204745
21	Anemia	15529	15522	14635	15745
22	Other hematological disease	20574	21627	20311	19232
23	Thyroid disorders	73936	73914	92746	82918
24	Diabetes	501251	452126	493689	503293
25	Other endocrine disorders	208082	188255	196280	198311
26	Vascular dementia	32318	36622	31832	32144
27	Drug addiction	3830	2921	3861	4167
28	Schizophrenia	49224	43562	48419	49136
29	Mood disorders	69727	72683	79177	81324
30	Neurosis	54524	45777	50019	48328
31	Mental retardation	7038	10398	7047	8501
32	Other psychiatric	6739	7501	7756	7936
33	Parkinson disease	21572	22190	22973	23039
35	Epilepsy	45056	35953	41097	41379
36	Cerebral Palsy	3041	3427	5369	5706
37	Autonomic nervous disorder	9538	5838	7874	8315
38	Other neurological disease	43170	43214	41381	39657
39	Conjunctivitis	102877	112781	127562	121895
40	Cataract	170778	162378	149433	156825
41	Refractory disorder	164194	153052	158107	166932
42	Other ophthalmic disease	231404	220523	228206	222683
43	Otitis externa	8383	6770	8106	8131
44	Other external ear disorders	14654	11350	10720	12271
45	Otitis media	35588	31191	27417	25392
46	Other middle ear diseases	6603	8220	7219	7486
47	Menier disease	7657	5593	6660	6240
48	Other inner ear diseases	759	759	759	759
49	Other ear diseases	14900	14086	16314	17298
50	Hypertension	936169	898315	939470	931321
51	Ischemic heart disease	179543	169507	189963	183444
52	Other heart diseases	142976	152981	166496	174881
53	Subarachnoid hemorrhage	3501	4677	3586	3972
54	Intracerebral hemorrhage	16267	13301	13806	14348
55	Cerebral infarction	212672	187581	211043	218356
56	Cerebral arteriosclerosis	5278	5492	5187	5343
57	Other cerebrovascular diseases	34703	36996	35315	35455
58	Atherosclerosis	35379	38096	41435	39485
59	Hemorrhoids	18118	21987	17350	19420
60	Hypotension	2368	2160	1686	1529

Table 1. Validation using Simulation Data (Continued)

Serial number	Diagnostic categories	right answer	PDM		MLE
			with arithmetic means	with PAE	
61	Other circulatory diseases	28444	27460	26674	28177
62	Acute nasopharyngitis (cold)	44160	49925	44762	46009
63	Acute tonsilitis	114234	114111	110615	112052
64	Other acute upper respiratory infections	214652	237450	245121	231606
65	Pneumonia	18515	20690	18021	15819
66	Acute bronchitis	157733	183161	164667	163924
67	Allergic rhinitis	81335	85051	80812	81402
68	Chronic sinusitis	57820	69922	67988	66234
69	Acute or chronic bronchitis	32834	32365	34241	35837
70	Chronic obstructive pulmonary disease	27780	21510	23540	27037
71	Asthma	217161	214210	226115	221951
72	Other respiratory diseases	41655	44884	46285	48766
76	Gastric and duodenal ulcer	200686	201830	201501	202535
77	Gastritis and duodenitis	165922	159634	173830	162164
78	Alcoholic liver disease	8531	2961	4390	5728
79	Chronic hepatitis (not alcohol related)	38963	41447	41131	41836
80	Cirrhosis (not alcohol related)	14723	10625	12742	11546
81	Other liver diseases	35452	31587	32540	33200
82	Cholelithiasis	16729	15165	19478	18849
83	Pancreatic disease	17628	16785	22081	23249
84	Other GI diseases	67276	58373	61586	58560
85	Skin diseases	21591	20859	17509	18209
86	Skin infection	213026	249688	189061	194844
87	Other skin diseases	75755	80844	79617	83621
88	Inflammatory polyarthropathies	88829	79476	90551	87834
89	Arthrosis	104190	127195	108002	110291
90	Spondylopathies	90204	120527	103554	96574
91	Intervertebral dis disorders	54575	59906	42678	39573
92	Cervicobrachial syndrome	23485	24381	9667	13734
93	Low back pain	49689	61607	43210	42435
94	Other vertebral diseases	27310	34257	27989	23506
95	Shoulder disorders	27632	29870	21444	22573
96	Osteoporosis	63514	60121	57063	56223
97	Other musculoskeletal disorders	68191	67112	72195	71770
98	Glomerular disease	27815	23624	26130	24450
99	Renal failure	728548	765301	687795	638235
100	Urolithiasis	20802	16864	18802	19841
101	Other urinary disease	61818	59363	54627	52778
102	Prostatic hypertrophy	53871	65042	62372	65234
103	Other male genital disorders	11747	10847	12554	10244
104	Menopausal disorders	29335	33804	35345	29763
105	Breast and female genital disorders	59646	53432	58198	58166
109	Other pregnancy related disorders	31620	25614	29662	29897
112	Congenital heart anomaly	7096	2543	3765	4052
113	Other congenital malformations	6921	5490	6218	6856
114	Symptomes and findings unclassified	75739	82203	85560	82382
115	Fracture	63089	72208	70769	70782
116	Head or abdominal njuries	4973	5425	5235	5257
117	Burn	6547	5594	5837	6176
118	Poisoning	2487	1612	1339	1455
119	Other external injury	152752	145123	147492	147854
	TOTAL	8334442	8334411	8334402	8269938
	SLOPE		0.9837	0.9862	0.9659
	INTERCEPT		1246.0	1057.3	2013.7
	R2		0.9926	0.9956	0.9935

PDM: Proportional Disease Magnitude PAE: Proportional Allotment Estimator
MLE: Maximum Likelihood Estimator

Table 2. Concurrent Validity using Actual Claims

Serial number	Diagnostic categories	PDM		MLE
		with arithmetic means	with PAE	
1	Intestinal infectious diseases	1225585	1621896	2297723
2	Tuberculosis	498774	474467	532847
3	Sexually Transmitted Diseases	346547	305498	389888
4	Viral infections with skin lesions	411500	621461	1018466
5	Viral hepatitis	53381	69120	−17828
6	Other viral disease	281539	212162	389345
7	Mycoses	900365	892107	1127977
8	Sequelae of infectious diseases	27040	0	−163073
9	Other infectious diseases	575241	389374	438925
10	Stomach cancer	664342	579100	1206757
11	Colon cancer	471453	533633	441892
12	Rectal cancer	211863	226900	145992
13	Liver cancer	358841	303546	254727
14	Lung cancer	678804	781548	1119435
15	Breast cancer	481016	681107	813398
16	Uterine cancer	72829	91291	66819
17	Malignant lymphoma	78769	124870	83699
18	Leukemia	21438	29924	23940
19	Other malignant neoplasms	2527689	2783415	3714578
20	Benign neoplasm	3815728	5047913	6438994
21	Anemia	3135726	1176414	983862
22	Other hematological disease	1056766	707157	960126
23	Thyroid disorders	1763021	1846030	2629737
24	Diabetes	22918	28301	−2602
25	Other endocrine disorders	6241898	6552406	14056624
26	Vascular dementia	189760	192018	208094
27	Drug addiction	70577	117960	106554
28	Schizophrenia	1752965	2578653	2797218
29	Mood disorders	1018147	1837882	2228521
30	Neurosis	1157413	1500621	1856245
31	Mental retardation	81229	42298	45542
32	Other psychiatric	784958	715578	958284
33	Parkinson disease	795621	691757	1030254
34	Alzheimer disease	77915	70949	103106
35	Epilepsy	780745	759194	1118589
36	Cerebral Palsy	218087	113609	91602
37	Autonomic nervous disorder	153200	160471	4318
38	Other neurological disease	933381	489947	362440
39	Conjunctivitis	2023180	1901823	1493792
40	Cataract	3259660	2787850	1473904
41	Refractory disorder	3482660	6954456	7870980
42	Other ophthalmic disease	3910830	2407838	1977178
43	Otitis externa	102728	163224	149140
44	Other external ear disorders	90077	142933	178170
45	Otitis media	130833	226904	338836
46	Other middle ear diseases	49690	114212	250785
47	Menier disease	127222	0	−61204
48	Other inner ear diseases	10500	0	−35565
49	Other ear diseases	264099	388483	590753
50	Hypertension	709366	1046858	1144591
51	Ischemic heart disease	2594268	1539316	789937
52	Other heart diseases	6066122	3810111	3450382
53	Subarachnoid hemorrhage	391363	484882	500644
54	Intracerebral hemorrhage	453568	502061	865783
55	Cerebral infarction	5532719	5846447	9243710
56	Cerebral arteriosclerosis	74768	71054	58478
57	Other cerebrovascular diseases	1406872	1571456	2404622
58	Atherosclerosis	1799707	558880	1207239

Table 2. Concurrent Validity using Actual Claims (Continued)

Serial number	Diagnostic categories	PDM		MLE
		with arithmetic means	with PAE	
59	Hemorrhoids	269757	552347	537377
60	Hypotension	319881	280143	32623
61	Other circulatory diseases	824617	909561	1186240
62	Acute nasopharyngitis (cold)	1057218	1506924	2199512
63	Acute tonsilitis	709017	994272	1356435
64	Other acute upper respiratory infections	2648874	3602926	4962389
65	Pneumonia	647835	441526	842183
66	Acute bronchitis	2973331	4491969	5744818
67	Allergic rhinitis	1838341	2246915	3442421
68	Chronic sinusitis	570705	628194	83185
69	Acute or chronic bronchitis	128516	238724	316240
70	Chronic obstructive pulmonary disease	2506918	1188044	1538186
71	Asthma	2489798	2781782	3562648
72	Other respiratory diseases	2378762	1859616	2250441
74	Gingivitis	5874	10190	6595
75	Other dental disorders	8882	12254	11022
76	Gastric and duodenal ulcer	4022918	4093314	4593668
77	Gastritis and duodenitis	5922871	5776411	6075661
78	Alcoholic liver disease	197655	132641	193408
79	Chronic hepatitis (not alcohol related)	918982	431132	156641
80	Cirrhosis (not alcohol related)	376383	252223	81147
81	Other liver diseases	932843	374792	121949
82	Cholelithiasis	1239348	865238	842172
83	Pancreatic disease	570421	449254	559539
84	Other GI diseases	6441859	3668146	2575853
85	Skin diseases	298086	505627	583989
86	Skin infection	2624170	2799504	4195819
87	Other skin diseases	2209889	2519770	4108901
88	Inflammatory polyarthropathies	1032028	1111215	2328836
89	Arthrosis	2690905	3712255	6380371
90	Spondylopathies	2307824	1949847	6014512
91	Intervertebral dis disorders	595584	937647	1694112
92	Cervicobrachial syndrome	528665	502278	－542586
93	Low back pain	420023	646869	548705
94	Other vertebral diseases	545868	697599	952192
95	Shoulder disorders	1297136	1314756	1691730
97	Other musculoskeletal disorders	2261039	2241600	4280709
98	Glomerular disease	806943	656659	479299
99	Renal failure	9332608	7162470	4531139
100	Urolithiasis	552953	370061	784601
101	Other urinary disease	4620076	3627988	3953495
102	Prostatic hypertrophy	2262849	2497411	3496668
103	Other male genital disorders	207924	281286	170310
104	Menopausal disorders	152956	259489	92531
105	Breast and female genital disorders	523955	756307	1003160
106	Abortion	47947	62467	82896
107	Toxemia	1621	1719	－2956
109	Other pregnancy related disorders	33821	56668	58794
110	Fetal growth disorder	18758	26549	23748
111	Other perinatal disorder	32483	37802	34404
112	Congenital heart anomaly	19428	43796	44499
113	Other congenital malformations	30994	43182	62122
114	Symptomes and findings unclassified	8875720	5131055	3391870
115	Fracture	1185814	1577619	2439301
116	Head or abdominal njuries	41367	32815	34459
117	Burn	55439	83892	87038
118	Poisoning	4869	6300	4851
119	Other external injury	3365688	4393837	7596788

Table 2. Concurrent Validity using Actual Claims (Continued)

Serial number	Diagnostic categories	PDM		MLE
		with arithmetic means	with PAE	
201	Hyperlipidemia	7097390	7018353	7424584
202	Hypertension not specified	12161650	18750628	22500982
203	Atopic dermatitis	183305	350536	351637
204	Arthrosis of knee	236354	127370	194269
205	Diabetes Mellitus	9380019	9206516	12494463
207	Diabetic nephropathy	1242811	592798	786471
208	Diabetic neuropathy	1066870	715920	1237721
209	Diabetic cataract	14621	0	13721
210	Diabetic retinopathy	1474469	1371957	771761
211	Hypertensive nephropathy	1440	0	−52232
213	Hemiplegia	91010	30535	−379981
214	Hepatitis C	725377	638274	1153384
215	Hepatocelular carcinoma	216452	147757	605715
216	NIDDM	28497	39867	128906
217	obesity	101513	9872	−46729
218	Exudative otitis media	238514	311925	443114
219	Amyotrophic Lateral Sclerosis	34	0	−15985
220	Spirocerebellar degeneration	166	0	−14848
221	Osteoporosis	1977963	1902301	1550282
222	Peripheral neuropathy	390493	226712	−98198
223	Fatty liver	885778	582109	310564
224	Lumbago	3904880	1867743	1994007
225	Hepatitis B	167599	129002	124949
226	Cervical cancer	120263	130105	237268
227	Endometrial cancer	21491	11576	−5055
228	Prostate cancer	1852575	1860812	1818445
229	IDDM	4464	5040	−1076
230	Allergic conjunctivitis	772360	928606	1003051
231	Essential hypertension	2886876	5995209	6764045
232	Angina pectoris	3600902	2544562	1626727
233	Acute Myocardial Infarction	106372	97814	135635
234	Carotid atherosclerosis	17257	0	−8342
235	Varix	97051	0	−37318
236	Influenza	1135212	1864523	2372903
237	Gout	471781	433979	412806
238	Spondylosis	1662012	2133028	990805
240	Cervical fracture	14837	23248	27534
241	Femoral fracture	1	0	−1054
	TOTAL	209754998	209754920	254448765

※serial numbers above 200 denote additional categories of Natori city
PDM: Proportional Disease Magnitude
MLE: Maximum Likelihood Estimator
PAE : Proportional Allotment Estimator

numerous zero disease-specific costs in actual claims are simply hard to accept.

We have established criteria for validation to be met for the methods to be suitable for claims analysis. Validity with simulation data is only a necessary condition and does not guarantee the satisfactory conditions: concurrent validity in actual claims. Here traditional estimation methods such as MRA and MLE failed to fulfill the satisfactory conditions although they fulfilled the necessary conditions. We believe this is the reason why claims analysis has long defied traditional estimation methods. MRA is meant to estimate the cost of a claim from diagnoses, not vice versa.

The distribution method, PDM, demonstrated a good concurrent validity with two different magnitudes but we believe even better magnitudes are possible. It is therefore necessary to continuously refine magnitude estimation for more valid and accurate claims analysis.

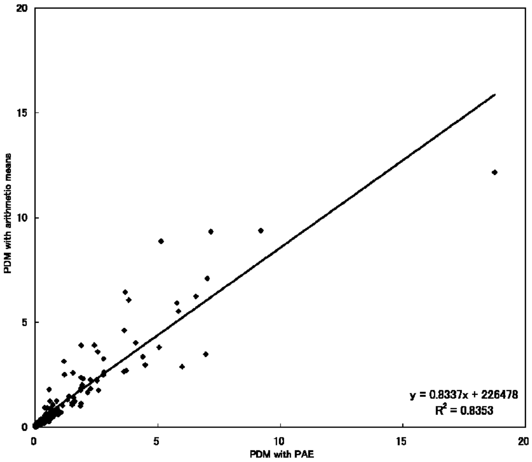

Figure 2. Correlation between PDM with arithmetic means and PDM with PAE
PDM using two different magnitudes estimated by arithmetic means with correction and Proportional Allotment Estimator (PAE)
Natori city outpatient claims (N=15,571, 209,754,920 yen)

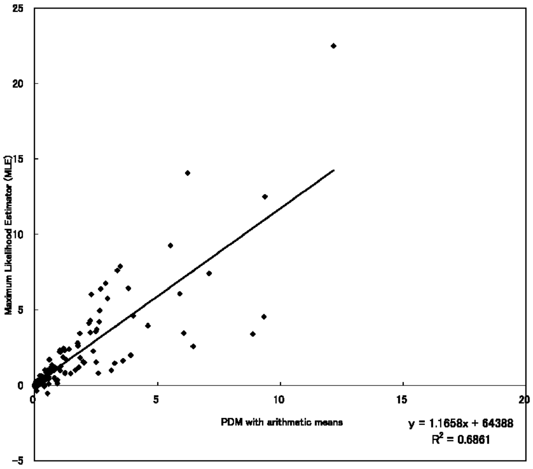

Figure 3. Correlation between MLE and PDM with arithmetic means
Note that estimates by MLE are negative in 18 categories
Natori city outpatient claims (N=15,571, 209,754,920 yen)

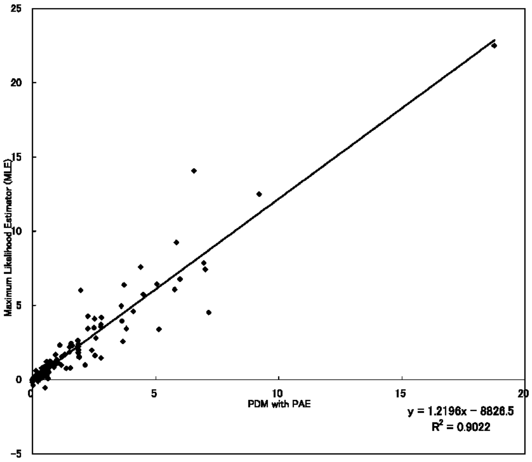

Figure 4. Correlation between MLE and PDM with PAE
Note that estimates by MLE are negative in 18 categories
Natori city outpatient claims (N=15,571, 209,754,920 yen)

Acknowledgement

This study was supported jointly by Grant-in-Aid for Scientific Research (C) "Development of data mining technique to detect drug adverse events from health insurance claims (theme No. 14570373, Principal investigator: Etsuji Okamoto)" and MHLW Health and Labour Sciences Research Grants, Research on Policy Planning and Evaluation "Methodological innovation and privacy protection for research use of health insurance claims (H15-Policy-016, Principal investigator: Yasuki Kobayashi)"

References

1) Okamoto E. Proportional Disease Magnitude [PDM] method for computerized health insurance claims. J Health Welfare Stat 1996; 43: 24–29.
2) Okamoto, E, Hata, E. Estimation of disease-specific costs in a dataset of health insurance claims and its validation using simulation data. Jpn J Public Health 2003; 50: 1135–1143.
3) Tango T. Can S-PLUS provide an artistic tool for statisticians?—A case of estimating parts from the whole. Proceeding from the 3rd Users' Conference of S-PLUS 2003: 1–15.
4) Tango describes the data presented in citation [3] as "health insurance claims" but we strongly question the authenticity. We presume that they were simulation data, which we provided to Tango, and not actual claims.
5) Okamoto originally called the method the "Proportional Disease Magnitude method" because it was intended for health insurance claims analysis. It is more appropriate to call it the "Proportional Distribution Method" and we now use the latter term. However, in this article we adhere to its original name.
6) Ministry of Health & Welfare Bureau of Health Insurance. Disease Classification for Social Insurance Statistics 〈119 classifications〉. Tokyo: Shakaihoken Jitsumu Kenkyujo, 1995.

7) Okamoto E, Hata E. Refinement of Proportional Distribution Method (PDM) with improved magnitude estimations and validation by Monte Carlo simulation. J Health Care Society [in press].

8) Ministry of Health, Labor & Welfare, Health Insurance Bureau. 1999 National Health Insurance Claims Survey. The Central Federation of National Health Insurance.

9) Ministry of Health, Labor & Welfare, Statistics & Information Bureau. 1999 Social Insurance Claims Survey. The Health & Welfare Statistics Association.

10) Streiner DL, Norman GR. Health Measurement Scales. 2nd edition. Oxford: Oxford University Press, 1995; 147–150.

11) Okamoto E. Reduction of influenza-related outpatient visits among community-dwelling elderly who received influenza vaccination. Jpn J Pharmacoepidemiol 2003; 8: 55–60.

(Received June 23, 2004; Accepted October 8, 2004)