Introducing fees for services with professional uncertainty

by Henk A. Flierman and Peter P. Groenewegen

A change in payment system of general practitioners from capitation to a mix of one-half capitation and one-half fee for service in Copenhagen, Denmark, resulted in a significant overall increase in diagnostic and curative services. The rate of increase differs between services. In this article, it is assumed that the rate of increase varies with doctors' professional uncertainty relative to the services studied. Professional uncertainty is measured as the degree to which performances of a service are determined by diagnoses made. The data validate the measure given the assumption.

Introduction

Since 1961, Denmark has had a system of payment by prospective estimation, by which about one-half of the income of the average general practitioner (GP) comes from capitation for patients on his or her list, and about one-half from fees for consultations and a set of mainly diagnostic and curative services. In Copenhagen City, this system was not introduced until October 1987. Until then, a payment system was in use in which GPs received the major part of their income from capitation. Copenhagen City, with about one-half million inhabitants, is the central part of the larger suburban area of Copenhagen County, with about 1 million inhabitants in all.

We examined the effect of the introduction of fees for services in October 1987, by comparing the change in number of services performed among Copenhagen City GPs with the change among GPs in the remainder of Copenhagen County. Copenhagen City GPs served as the experimental group, and GPs in the remainder of Copenhagen County, where the mixed system of GP remuneration already existed, served as the control group. We compared numbers performed of a sample of the diagnostic and curative services newly paid for in Copenhagen City, comparing numbers per week per 1,000 registered patients 6 months before October 1987 with numbers 6 months after that, and 1 year after that.

Total numbers of both sampled diagnostic and sampled curative services showed a significantly larger increase among Copenhagen City GPs than among GPs in the remainder of Copenhagen County. Numbers both 6 months and 1 year after October 1987 were compared with numbers 6 months before that. These results were stable in that they were obtained equally with differing statistics (Krasnik et al., 1990; Flierman and Groenewegen, 1991).

We also investigated the changes in single diagnostic and curative services and concluded that the introduction of fees produced far larger increases in some services than in others. Here we ask how these differences can be explained.

In health economics, increases like the ones mentioned are explained by stating that income is an argument in GPs' utility function. Yet income maximization is considered to be constrained by professional medical standards on when to perform a service, as another argument in that function. The strength of that constraint, however, is considered to vary with consensus on standards (Evans, 1984).

In health services research, variation in consensus on when to perform services is nowadays attributed to varying strength of scientific foundation of standards, as perceived by the medical community (Wennberg, Barnes, and Zubkoff, 1982; Wennberg, 1987; Wolff, 1989). The weaker the perceived scientific foundation of a standard on when to perform a service, the greater is the lack of consensus, or professional uncertainty relative to it. And the greater the uncertainty, the larger the variation in performance of a service. Comparing surgical interventions performed in hospitals, differences between hospital markets in numbers per capita that are attributed to varying degrees of professional uncertainty range from no differences up to tenfold differences.[1]

Attribution of variation to professional uncertainty results from standardizing rates for age and sex distributions of populations, and excluding random variation under a Poisson assumption (McPherson et al., 1982). The magnitude of the remaining systematic variation is then taken as a measure of professional uncertainty (McPherson, 1989).

In a later section, another measure of professional uncertainty will be introduced that differs in content as well as in form from the one previously described. As for content, it takes account of morbidity itself, instead of using age and sex as proxies. As for form, it measures uncertainty in terms of indetermination rather than variation. The concept to be measured is uncertainty among providers about when to perform a service. Therefore, the degree to which performances of a service are determined by assessed (morbidity) conditions reflects the structure of the concept more explicitly. As we will argue at the end of "Measuring professional uncertainty," with sufficiently precise diagnostic data the components of the measure can be directly compared with medical knowledge to examine content validity.

We will also discuss the changes in single diagnostic and curative services in Copenhagen City when fees for them were introduced, comparing them with the changes in Copenhagen County

Reprint requests: Peter P. Groenewegen, Netherlands Institute of Primary Health Care, P.O. Box 1568, 3500 BN Utrecht, The Netherlands.

[1] See Diehr et al. (1990) for a discussion of the statistical significance of extremal quotients like "tenfold."

(Flierman and Groenewegen, 1991). Next, we will use the relative changes in Copenhagen City to examine the construct validity of our measure of professional uncertainty. We will do so by investigating the health economics prediction previously discussed. The prediction says that the introduction of a fee for a service stimulates its performance more when professional uncertainty concerning it is greater.

Data and method

Changes in services performed

Comparing March 1987 with March 1988 data (Flierman and Groenewegen, 1991), increases in diagnostic and curative services were shown to be larger in Copenhagen City than in Copenhagen County on a 5-percent significance level. Comparing March 1987 with November 1988 data, they were on a 1-percent significance level. Hence in the latter comparison the evidence is strongest that services in general are performed more often because fees for them have been introduced. Therefore, the March 1987 and November 1988 comparison will be used here to investigate the health economics prediction on varying degrees of increase.

In Copenhagen County, fees were already paid for 43 diagnostic and 27 curative services. Our data on the Copenhagen County control group come from claims files on these services performed by all 326 and 329 GPs practicing in Copenhagen County in March 1987 and November 1988, respectively. These data are available on the aggregated level of regional groups of doctors only. Copenhagen County has 8 such regions with a minimum of 10 and a maximum of 82 doctors per region.

Our Copenhagen City experimental group is a sample of 72 out of 265 GPs practicing there in March 1987. These doctors were mainly self-selected by willingness to volunteer in our recording survey (Krasnik et al., 1990). They recorded their consultations for 1 week each in the months concerned. The recording form included a sample of 13 out of the 43 diagnostic services for which fees were introduced in Copenhagen City as well, and a sample of 8 out of the 27 curative services for which fees were introduced. Services were included in the sample that were frequently performed in Copenhagen County, and for which the average Copenhagen City GP would need no new equipment. Effects of the introduction of fees would therefore not be curbed by postponed investments.

Total numbers entering the comparison are reported in Table 1. Numbers of registered patients in the county in March 1987 were no longer available at the time of data collection. Numbers of patients in March 1988 are used as an estimate. In March 1988, 329 doctors were practicing in the county. Numbers of doctors per region in March 1988 are also applied to the March 1987 and November 1988 comparison.

The numbers to be compared are numbers per week per 1,000 registered patients. For the sample of city doctors, these numbers are assessed per doctor. For the county, they are assessed per region. County averages are weighted by numbers of doctors per region.

In testing the difference between the changes in city and county, we compare proportional changes. Proportional changes are argued to be a better measure than sizes of change (Flierman and Groenewegen, 1991). Because the county data are not available per doctor, we assume that the variance of changes among county doctors equals the one among city doctors. Under that assumption:

$$t = ((chge_{meas.city} - chge_{exp.city})/sd_{chge.meas.city}) * * (n_{city} * n_{county})/(n_{city} + n_{county}),$$

where

- $chge_{meas.city}$ is the measured average change in the city,
- $chge_{exp.city}$ is the expected average change in the city, assessed as (proportional average change in the county) * (initial value in the city),
- $sd_{chge.meas.city}$ is the standard deviation of the measured changes in the city, and
- n_{city} and n_{county} are the numbers of doctors in city and county.

The *p*-values associated with this *t*-statistic are reported in Table 5.

Diagnoses made

Diagnoses made are not in the billing data of the Copenhagen County doctors in the control group. The recording form used by the experimental group in Copenhagen City included a list of complaints and

Table 1

Numbers of general practitioners, patients, consultations, and services in the natural experiment in Copenhagen City and County

Data category	March 1987		November 1988	
	City (week)	County (month)	City (week)	County (month)
General practitioners	72	326	72	329
Registered patients	123,964	[1]512,045	127,009	[1]512,045
Consultations	9,516	194,984	10,419	257,827
Diagnostic services (13)	525	24,622	892	33,843
Curative services (8)	106	6,029	199	8,857

[1]An estimate, using number of patients in March 1988.

NOTE: Copenhagen City represents the experimental group whereas the County represents the control group.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

diagnoses derived from the International Classification of Health Problems in Primary Care—second revision (ICHPPC-2) (Classification Committee, 1983). In measuring professional uncertainty among Copenhagen City doctors in the next section, we will relate services performed to diagnoses made according to that list.

Table 1 showed that Copenhagen City doctors performed the sampled diagnostic and curative services nearly twice as often in November 1988 as they did in March 1987. Where services are performed more often, the degree to which their performance is determined by diagnoses made can be better assessed. Therefore, the November 1988 data will be used to assess these degrees of determination.

Formally, in doing so we weaken our test of the predicted effect of professional uncertainty. A greater professional uncertainty among doctors as measured after the change in payment system may not be a modifier of the effect of the introduction of fees, but another effect of it. It may be that fees tend to increase the numbers of services by increasing the range of health conditions in which they are performed. And it may be that services vary in their openness to such a tendency. Then, in varying degrees, "more servicing, which may initially be a response to economic factors, becomes over time the new standard" (Evans, 1984). Our analysis rests on the assumption that 1 year is too short a time span for that. We introduce that assumption in order to avoid small numbers.

We further avoid small numbers by including data on four Copenhagen City GPs who are not in the experimental group because they did not record their consultations in March 1987, but who did do so in November 1988. Uncertainty among GPs about when to perform a service is therefore assessed for 76 GPs.

More than one diagnosis could be recorded in the list of complaints and diagnoses mentioned. The numbers of consultations with differing numbers of diagnoses are listed in Table 2. The large number of consultations with no diagnosis partly reflects the fact that in general practice a diagnosis is not always made. It may be unclear what diagnosis should be made, or the reason for the consultation may be of an administrative nature (e.g., getting the result of a diagnostic test).

Partly, the large number of consultations without any diagnosis is also an artifact of data collection. Our recording form contained a list of 34 complaints and diagnoses, with the request either to pick an item on that list or else to write down the diagnosis. The written diagnoses most closely related to one another and largest in number were grouped into 11 extra items afterwards; many diagnoses were lost both in collecting and in processing the data.

Table 2

Numbers of consultations with 0 through 5 diagnoses as coded from records of 76 Copenhagen City general practitioners: November 1988

Diagnoses	Consultations	Percent
Total	10,947	100
None	4,001	37
1	6,239	57
2	614	6
3	73	1
4	17	0
5	3	0

NOTE: Percent totals exceed 100 because of rounding.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

With no diagnosis coded, the degree to which the performance of a service is determined by the diagnosis cannot be assessed; with more than one diagnosis, coded complexities of comorbidity are introduced with hardly any increase in numbers of consultations to be analyzed (Table 2). Therefore, uncertainty among GPs about when to perform a service is assessed with data on 6,239 one-diagnosis consultations.

In Table 3, the diagnoses are listed that were made in these consultations and, moreover, were made five times or more while a service in our sample was performed. The diagnoses listed are derived from ICHPPC-2 by selection and combination. Next, we grouped them into umbrella categories, like "acute infections." Diagnoses are not listed if they were not made when a service was performed, or if there is a suspicion of measurement error.

The combinations of diagnoses and services have been checked on their medical plausibility. Combinations that were clearly produced by measurement error never arose in more than four consultations. Therefore, all combinations that arose in fewer than five consultations were excluded. Hence, only diagnoses coinciding with services in at least five consultations are listed here.

Diagnoses not listed are put together in the residual categories, such as "other acute infections," and in the umbrella category "psychiatric and neurological disorders." No psychiatric or neurological disorder was ever diagnosed when a service in our sample was performed (numbers under suspicion of measurement error excluded).

The 18 diagnoses listed (residual categories excluded) can be used in measuring professional uncertainty on when to perform the 21 services in our sample. However, for diagnoses and services, similar restrictions hold. Not all 21 services were performed in combination with one of the diagnoses listed in sufficient numbers to be above suspicion of measurement error.

For this reason, nine services are omitted in the next section. These are: taking a blood sample, proctoscopy, electrocardiogram, erythrocytes sedimentation rate (ESR) measurement, urine microscopy, removing ear wax, removing corpora aliena (with two specifications), and bladder catheterization. A 10th service, performing a pregnancy test, was omitted because its most obvious reason is not included in either ICHPPC-2 or our recording form—a woman's question of whether or not she is pregnant.

For the remaining 11 services, professional uncertainty on when to perform them, in terms of the 18 diagnoses listed, could be measured.

Table 3

Numbers of November 1988 one-diagnosis consultations of 76 Copenhagen City general practitioners in which any sampled service was performed at least five times, by diagnosis made

Diagnosis	Number	Percent
Total	6,239	100
Acute infections		
Total	1,536	25
Upper respiratory infection (head cold)	521	34
Acute tonsillitis	139	9
Urinary infection	209	14
Conjunctivitis	117	8
Other acute infections	550	36
Acute injuries		
Open wound	37	1
Chronic disorders		
Total	1,189	19
Hypertension	334	28
Diabetes mellitus	98	8
Anemia	55	5
Other chronic disorders	702	59
Psychiatric and neurological disorders	1,219	20
Gynecological and obstetric disorders		
Total	625	10
Menstrual disorders	164	26
Fluor vaginalis	193	31
Menopause symptoms	69	11
Other gynecological disorders	69	11
Obstetric disorders	130	21
Joint and muscle disorders		
Total	975	16
Non-degenerative joint pains	258	26
Other joint and muscle disorders	717	74
Dermatologic disorders		
Total	658	11
Warts/condylomata acuminata	122	19
Dermatitis/eczema/rash	366	56
Abscesses and other localized inflammations	109	17
Benign tumor	61	9

NOTES: Percent totals may not equal 100 because of rounding. Totals under the boldface umbrella categories add up to 100 percent. Percentages of single diagnostic categories add up to 100 percent within umbrella categories.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

Measuring professional uncertainty

As we previously discussed, professional uncertainty is measured here in terms of indetermination of performances of a service by morbidity conditions assessed.

With full professional certainty among doctors, there would be full consensus on standards on when to perform a service. Such a standard would include a full description of the conditions in which a service should be performed. Full consensus on such a standard would show when every doctor always performed the service upon assessing a condition described by the standard, with no doctor ever performing the service in other conditions.

In such a state of affairs, the number of times any doctor performed a service would equal the number of times he or she assessed the relevant conditions.

Now assume that the ideal state of affairs is still there, but conditions are not described fully in the data. For example, every doctor always removes ear wax with hearing problems and when the external ear is infected. No doctor ever removes ear wax in other conditions. From the data we know of hearing problems as such, but we know only of infections of the ear generally. Next, assume that the relevant but unknown particulars within the conditions that we do know of are equally distributed in every doctor's practice. To continue the example, assume that in every doctor's practice one-half of the ear infections are in the external ear. In that case, every doctor would remove ear wax a number of times equal to the doctor's assessments of hearing problems plus one-half of the assessments of ear infections.

In a regression over all doctors, the non-standardized coefficient (b) of ear wax removals on hearing problems would be 1, and on ear infections it would be ½. Still, the coefficient of determination (R^2) would be 1, expressing full consensus among doctors on when to remove ear wax.

Thus, if the assumptions are valid, an empirical value of R^2 would be an appropriate measure of the degree to which a real state of affairs approximates the ideal one: Then R^2 is an appropriate measure of professional certainty among doctors.

It seems safe to assume that the relevant conditions are not fully described by the diagnoses listed for any of the services (Table 3).

By the same token, there is more hazard in the assumption of equal distributions of relevant particulars. A priori, the equal distributions assumption is more valid when the conditions on which we do have data are more specific themselves. On the face of it, the ones in Table 3 are quite unspecific indeed.

Hence, when the diagnoses listed in the previous section enter the measurement of professional uncertainty by R^2, their global character makes content validity questionable a priori. However, the test of the health economics prediction will decide whether content validity is good enough for construct validity to show.

In Table 4, R^2 is determined for 11 services. The numeric content is best described by taking one service as an example. For this we chose the service listed first, taking a cervical smear.

Cervical smears are taken with three diagnoses: menstrual disorders, fluor vaginalis, and menopause symptoms. These three diagnoses are made in 164, 193, and 69 consultations respectively; these numbers of consultations from Table 3 are repeated in Table 4. Because our total number of one-diagnosis consultations is 6,239, some other diagnosis is made in 5,813 consultations. Also described are different ways to what degree services are rendered when diagnoses are made.

Table 4

Numbers of consultations and performances of a service by diagnosis, with dependence[1] of performances on diagnoses made, for 11 out of 21 sampled services performed by 76 Copenhagen City general practitioners: November 1988

Diagnosis	Procedure		
	Consultation	Performance	Dependence
Cervical smear			
Menstrual disorders	164	15	+.24
Fluor vaginalis	193	26	+.17
Menopause symptoms	69	6	+.32
Other diagnoses	5,813	16	—
R^2 = .48			
Hemoglobin measurement			
Diabetes mellitus	98	5	+.12
Anemia	55	16	+.67
Obstetric disorders	130	5	+.06
Other diagnoses	5,956	33	—
R^2 = .25			
Blood glucose (photometer)			
Diabetes mellitus	98	14	+.14
Other diagnoses	6,141	7	—
R^2 = .10			
Streptoculture or urine culture			
Upper respiratory infection (head cold)	521	6	+.05
Acute tonsillitis	139	5	+.09
Urinary infection	209	18	+.11
Other diagnoses	5,370	12	—
R^2 = .06			
Inoculation for cultivation			
Upper respiratory infection (head cold)	521	11	−.02
Acute tonsillitis	139	14	+.39
Urinary infection	209	8	+.03
Conjunctivitis	117	7	−.07
Menstrual disorders	164	5	+.02
Fluor vaginalis	193	46	+.47
Other gynecological disorders	69	5	+.23
Abscesses and other localized inflammations	109	5	+.53
Other diagnoses	4,718	19	—
R^2 = .49			
Urine test with stick			
Urinary infection	209	63	+.53
Hypertension	334	7	+.00
Diabetes mellitus	98	19	+.35
Fluor vaginalis	193	10	−.08
Obstetric disorders	130	6	+.07
Other diagnoses	5,275	29	—
R^2 = .31			
Urine culture with sensitivity			
Urinary infection	209	19	+.11
Other diagnoses	6,030	2	—
R^2 = .14			
Removing warts			
Warts/condylomata acuminata	122	43	+.35
Other diagnoses	6,117	3	—
R^2 = .33			
Incision or excision of abscess or tumor			
Abscesses and other localized inflammations	109	12	+.05
Benign tumor	61	7	+.08
Other diagnoses	6,069	7	—
R^2 = .02			
Treating a large wound			
Open wound	37	11	+.35
Other diagnoses	6,202	4	—
R^2 = .44			
Dressing an immobilizing bandage			
Non-degenerative joint pains	258	8	+.03
Other diagnoses	5,981	9	—
R^2 = .01			

[1]Non-standardized regression coefficients *b* (with coefficients of determination R^2-adjusted), after aggregation into numbers per doctor.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

Cervical smears shows the number of times that the service is rendered when each of the three diagnoses is made, and when some other diagnosis is made. Thus, in 164 consultations menstrual disorders are diagnosed, and a cervical smear is taken 15 times.[2] In the 5,813 consultations in which anything different from menstrual disorders, fluor vaginalis, or menopause symptoms is diagnosed, a cervical smear is taken 16 times in all. Yet with no other diagnosis taken separately, a cervical smear is made more than four times, and thus above suspicion of measurement error. The numbers discussed so far are stepping stones in assessing professional uncertainty among the 76 Copenhagen City doctors about when to take a cervical smear.

[2]For an American reader, this and other proportions in the table may stand for a surprisingly low rate of performance. Relevant differences are that patients need a referral by their GP to visit a medical specialist, and that suing a doctor for malpractice is highly unusual, so doctors can afford to practice the art of "masterful inactivity," of waiting and seeing.

The results of a regression analysis are reported in which GPs are the units of analysis. The dependent variable is the number of times each GP makes a cervical smear. The independent variables are the numbers of times each GP diagnoses menstrual disorders, fluor vaginalis, and menopause symptoms. The number of times other diagnoses are made does not enter the regression equation as an independent variable. Yet the 16 times that a cervical smear was made with other diagnoses are included in the values of the dependent variable.

With this reservation, the empirical counterpart of our ideal state example is shown for cervical smears. One reads that, on average, GPs take a cervical smear 1 out of 4 times that they assess menstrual disorders: The non-standardized regression coefficient is + .24. Yet counted over all GPs, they take a cervical smear fewer than 1 out of 10 times that they assess menstrual disorders—15 out of 164 times. The difference disappears when a service is performed with "other diagnoses," as is shown with "urine culture with sensitivity" and "removing warts." Hence, the difference stems from excluding "other diagnoses" while including services performed with these in the total number of performances entering the regression equation. This enlarges the values of the non-standardized regression coefficients.

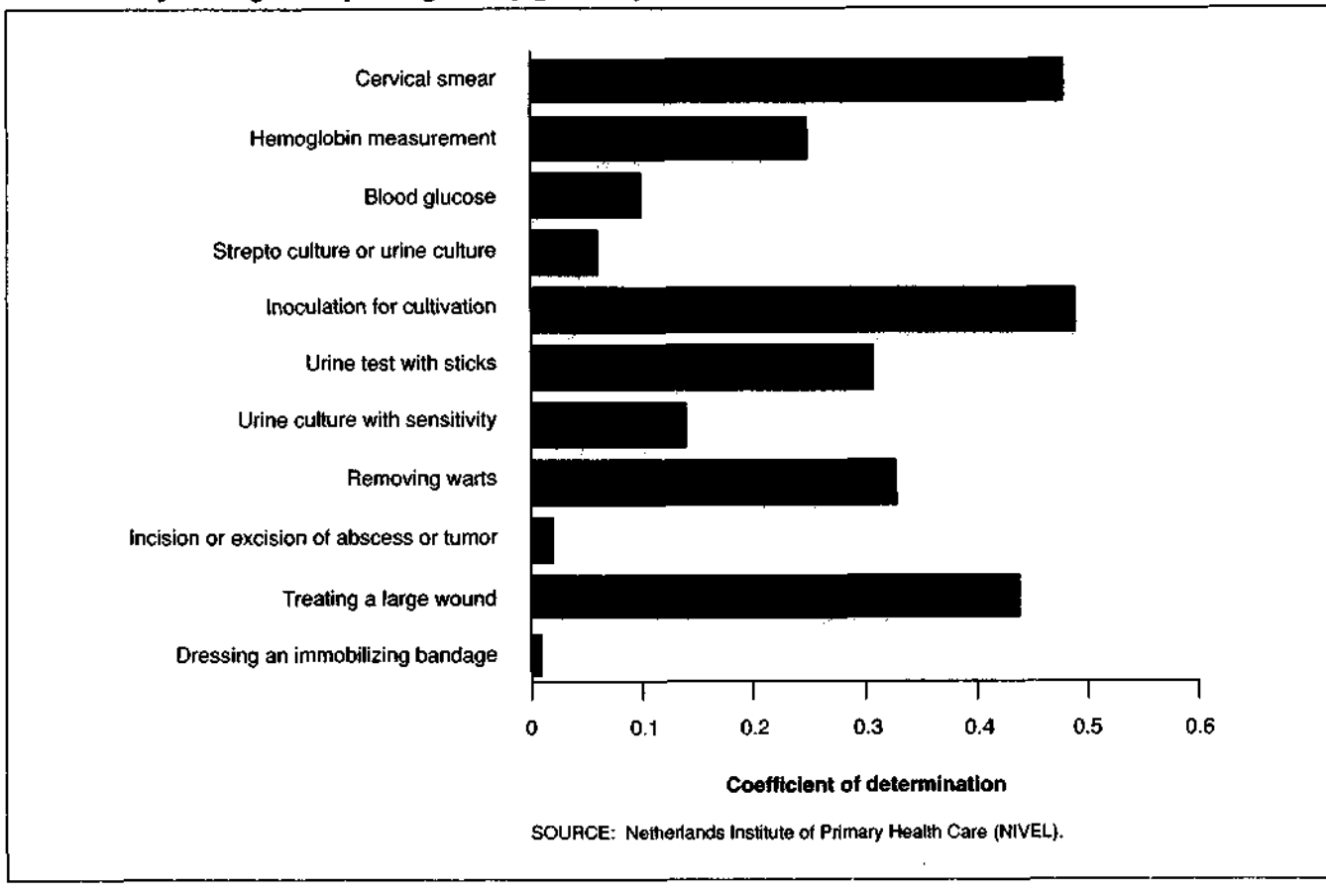
As another effect, it also reduces the value of R^2 when a service is often performed with other diagnoses. Performances of a service with diagnoses other than those entering the regression equation cannot be explained in the equation. But then again, when performances of a service are heavily scattered over other diagnoses, this can be taken as a secondary indication of indetermination; that is, if measurement errors are in the recording of diagnoses, and if these errors do not vary in number with the service performed. Under these assumptions, R^2 should be lower for heavily scattered services.

We could have avoided both effects by excluding the performances of the service with other diagnoses from the values of the dependent variable. But then the secondary indication of indetermination previously discussed would not have been covered by R^2, and the measure would have lost empirical coverage.

Given the enlarged regression coefficients, one reads that on average GPs take a cervical smear 1 out of 4 times that they assess menstrual disorders. Likewise, they do so 1 out of 6 times with fluor vaginalis (+ .17), and 1 out of 3 times when menopause symptoms are assessed (+ .32). Deviations from these averages indicate professional uncertainty as we introduced it in our ideal state example, under the equal distributions

Figure 1

Certainty among 76 Copenhagen City general practitioners relative to 11 services: November 1988



SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

assumption. These deviations and the scattered performances of the service taken together let R^2 sink from 1 to .48.

As measured in Table 4, the degree of professional certainty about when to perform the services listed varies from .01 for dressing an immobilizing bandage to .49 for an inoculation for cultivation. This is summarized visually in Figure 1.

One would like to judge the content validity of this measure. In principle, one could do so by examining whether the reported dependences of services performed on diagnoses made reflect medical knowledge. One could proceed by neglecting non-significant coefficients, taking a significant positive coefficient as evidence of agreement that a diagnosis is a good reason to perform a service, and taking a significant negative coefficient to show agreement that a diagnosis is a bad reason to perform the service.

In the latter instance, the finding would be that although a service is sometimes performed with a disorder, doctors who see the disorder more often, perform the service less often, other things being equal. This would reflect medical knowledge if experience shows (to the experienced) that the service generally does not help or give a definitive answer with the disorder at hand.

In Flierman, Groenewegen, and Stokx (1991), we applied the measure presented here to more precise diagnostic information, and could indeed show content validity with few exceptions, following the rules previously described.

However, owing to the global character of the diagnostic categories used in Table 4, here one would inevitably run into speculation on the precise nature of complaints and diagnoses. Therefore, here it seems more fruitful to examine whether content validity is good enough to produce construct validity, as described later.

Testing for construct validity

To measure professional uncertainty among doctors concerning services, data were analyzed on one-diagnosis consultations of 76 Copenhagen City doctors in November 1988. Now we consider the R^2 values obtained as characteristics of services, and assume that they are not affected yet by the change in payment system in Copenhagen City in October 1987. We relate them to changes in performance of services in all consultations by 72 Copenhagen City doctors between March 1987 and November 1988. The changes examined are relative to those among Copenhagen County GPs, and can therefore be attributed to the change in payment system in Copenhagen City.

R^2 values were obtained for 11 out of 21 services. In Table 5, changes in performance among Copenhagen City and Copenhagen County doctors are reported for 19 of them. ESR measurement and bladder catheterization are omitted. The 72 Copenhagen City GPs did not measure ESR at all in their week of March 1987, and did not catheterize any bladder in their week of November 1988. Hence, for ESR measurement, no proportional increase can be assessed, and attributing a 100-percent decrease (from .02 to 0.00 per 1,000) to bladder catheterization seems to be overstretching the data.

Among these 19 services, 18 tend to increase in number more strongly in Copenhagen City than they do in Copenhagen County. For 8 among the 18, the difference is significant on at least a 5-percent level. The relative increases vary from 5 percent for taking a cervical smear to 585 percent for blood glucose measurement, and of the significant ones the lowest relative increase is 51 percent for urine tests with sticks. The question to be answered here is how these varying rates of relative increase can be explained.

More specifically, the question is whether they can be explained from varying degrees of professional uncertainty among GPs about when to perform them. The prediction to be tested is: Services about which GPs are less certain professionally show a higher relative rate of increase when fees for them are introduced.

For 11 services, this prediction is tested by rank order correlation in Table 6. On a 5-percent significance level and among all 11 services, those about which doctors are less certain professionally as measured by a lower R^2 show a larger relative proportional increase. Among the eight services whose relative increase is significant on a 5-percent level, the association is not significant anymore.

What is to be concluded from these opposing results? The most relevant circumstance seems to be that in a serial testing of differences like the one reported in Table 5, some significant results can be expected to show up by chance. Therefore, the significance of differences is a less decisive proof that differences between changes are real and that relative changes are caused by the change in payment system than the simple fact that all relative changes in Table 6 are positive. The probability of such an outcome under chance is $(\frac{1}{2})^{11} = .05$ percent, and the probability under chance of the outcome in Table 5 is $19*(\frac{1}{2})^{19} = .004$ percent. We therefore consider the association among all 11 services in Table 6 to be the more decisive test of the construct validity of our measure. By that test, construct validity is confirmed.

Discussion

As discussed in an earlier section of this article, content validity of the measure of professional uncertainty introduced was questionable a priori owing to the global character of the diagnostic categories used. We examined whether content validity was good enough to produce construct validity. Indeed, it did: The health economics prediction introduced in the first section of this article was confirmed in the previous section.

The globality of diagnostic categories and the frequency of measurement errors are the main limitations of our study. Globality made it impossible to judge content validity. Because of measurement

Table 5

Changes in numbers of services per week per 1,000 registered patients performed, by 72 Copenhagen City general practitioners and 329 Copenhagen County general practitioners for 19 out of 21 sampled services

Physician services	Location	March 1987	November 1988	Proportional changes	Relative change
Diagnostic				Percent	
Blood sample	City	.32	.38	+22	+18
	County	.56	.58	+4	
Cervical smear	City	.96	1.15	+20	+5
	County	1.44	1.65	+15	
Pregnancy test	City	.26	.31	+19	+11
	Coutny	.48	.52	+8	
Proctoscopy	City	.02	.07	+332	+317
	County	.16	.18	+15	
Electrocardiogram	City	.01	.02	+108	+109
	County	.31	.31	−1	
Hemoglobin measurement	City	.54	.83	+54	*+52
	County	1.16	1.18	+2	
Blood glucose (photometer)	City	.04	.29	+633	***+585
	County	.28	.41	+48	
Streptoculture or urine culture	City	.14	.44	+223	*+211
	County	1.99	2.22	+12	
Inoculation for cultivation	City	.72	1.38	+91	*+68
	County	.72	.89	+23	
Urine test with sticks	City	1.15	1.85	+61	**+51
	County	2.77	3.05	+10	
Urine microscopy	City	.06	.10	+57	+53
	County	.89	.93	+4	
Urine culture with sensitivity	City	.05	.20	+275	+265
	County	.91	1.01	+10	
Curative					
Removing warts	City	.17	.44	+162	*+114
	County	.82	1.21	+48	
Removing ear wax	City	.40	.47	+17	+16
	County	.84	.86	+1	
Removing corpora aliena from eye/ear/nose/throat	City	.02	.01	−62	−41
	County	.09	.07	−21	
Removing corpora aliena from skin/from under nail	City	.05	.07	+41	+54
	County	.22	.19	−13	
Incision or excision of abscess or tumor	City	.08	.26	+221	**+198
	County	.36	.44	+23	
Treating a large wound	City	.07	.13	+75	+66
	County	.19	.21	+9	
Dressing an immobilizing bandage	City	.08	.24	+194	*+173
	County	.40	.48	+21	

*Statistically significant at $p < .05$ level.
**Statistically significant at $p < .01$ level.
***Statistically significant at $p < .001$ level.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

errors, we had to introduce extra assumptions in order to avoid loss of empirical coverage of our measure.

Both limitations can be overcome when complaints and diagnoses are written down by GPs as they assess them, and are centrally coded into a sufficiently precise classification by medically trained personnel. Thus the data were collected that were analyzed in Flierman, Groenewegen, and Stokx (1991). The classification used there is the 1989 NIVEL revision of the International Classification of Primary Care (Lamberts and Wood, 1987). With these data, content validity of our measure could be shown with few exceptions.

Further research should assess content validity of the measure introduced relative to any adequate diagnostic classification. The health economics prediction investigated here should be tested in natural experiment

Table 6

Professional certainty relative to 11 sampled services among 76 Copenhagen City general practitioners and relative proportional change in the performance of these services, by 72 of these general practitioners

Physician services	Professional certainty R^2	Relative change (percent)
Inoculation for cultivation	.49	*+68
Cervical smear	.48	+5
First treatment of large wound	.44	+66
Removing warts	.33	*+114
Urine test with sticks	.31	*+51
Hemoglobin measurement	.25	*+52
Urine culture with sensitivity	.14	+265
Blood glucose (photometer)	.10	*+585
Streptoculture or urine culture	.06	*+211
Incision or excision of abscess or tumor	.02	*+198
Dressing an immobilizing bandage	.01	*+173

	Rank order correlation	Significance
All 11 services	-.65	.032
8 services with significant relative change	-.57	.139

*Statistically significant at .05 level.

SOURCE: Netherlands Institute of Primary Health Care (NIVEL).

conditions like the one in Copenhagen, Denmark in 1987, using any diagnostic classification relative to which content validity of the measure has been demonstrated.

Confirmation of the prediction would then support the assertion that doctors react to the introduction of fee for service as income maximizers under the constraint of professional standards on when to perform a service, while the strength of that constraint varies with professional certainty about standards.

For health politics, such results would imply that, if the introduction of fees for services is considered, services should be chosen for which constraints on income maximization are strong, and constraints should be strengthened further.

Thus, services should be chosen with a strong scientific foundation of standards on when to perform them, as perceived by the medical community. That constraint could be strengthened further by developing and publishing standards, and by stimulating adherence to them through postgraduate education and peer review.

Acknowledgment

The study is part of a larger project carried out by a Danish-Dutch research group consisting of the authors and Allan Krasnik, Peter van Scholten, Mogens Trab Damsgaard, Poul A. Pedersen, Gavin Mooney, and Adam Gottschau. The authors gratefully acknowledge the financial support of the Dutch Commissie Programma Evaluatie.

References

Classification Committee of WONCA (World Organization of National Colleges, Academies, and Academic Associations of General Practitioners/Family Physicians): ICHPPC-2-Defined. *International Classification of Health Problems in Primary Care. 3rd edition.* Oxford. Oxford University Press, 1983.

Diehr, P., Cain, K., Connell, F., and Volinn, E.: What Is Too Much Variation? The Null Hypothesis in Small-area Analysis. *Health Services Research* 24(6):741-771, 1990.

Evans, R.G.: *Strained Mercy: The Economics of Canadian Health Care.* Toronto. Butterworth and Company, 1984.

Flierman, H.A., and Groenewegen, P.P.: *Effects of Introducing Fees for Services in Copenhagen, Denmark.* Utrecht, The Netherlands NIVEL, 1991.

Flierman, H.A., Groenewegen, P.P., and Stokx, L.J.: *Assessing the Diagnosis-determinedness of Services among Dutch General Practitioners.* Utrecht, The Netherlands. NIVEL, 1991.

Krasnik, A., Groenewegen, P.P., Pedersen, P.A., et al.: Changing Remuneration Systems: Effects on Activity in General Practice. *British Medical Journal* 300(6741):1698-1701, 1990.

Lamberts, H., and Wood, M.: *ICPC International Classification of Primary Care.* Oxford. Oxford University Press, 1987.

McPherson, K.: International differences in medical care practices. *Health Care Financing Review,* 1989 Annual Supplement: 9-20. HCFA Pub. No. 03291. Office of Research and Demonstrations, Health Care Financing Administration. Washington. U.S. Government Printing Office, Dec. 1989.

McPherson, K., Wennberg, J.E., Hovind, O.B., and Clifford, P.: Small-area Variations in the Use of Common Surgical Procedures: An international comparison of New England, England, and Norway. *New England Journal of Medicine* 307(21):1310-1314, 1982.

Wennberg, J.E.: The Paradox of Appropriate Care. *Journal of the American Medical Association* 258(18):2568-2569, 1987.

Wennberg, J.E., Barnes, B.A., and Zubkoff, M.: Professional Uncertainty and the Problem of Supplier-Induced Demand. *Social Science and Medicine* 16(7):811-824, 1982.

Wolff, N.: Professional Uncertainty and Physician Medical Decision-making in a Multiple Treatment Framework. *Social Science and Medicine* 28(2):99-107, 1989.