

Review of Physician Outcome Measurement Analysis

**Prepared for:
Ohio Bureau of Workers' Compensation**

**Marjorie L. Baldwin, PhD
W. P. Carey School of Business
School of Health Management and Policy
Arizona State University**

**William Lohman, MD
School of Public Health
University of Minnesota**

October 1, 2004

1. Introduction

The Ohio Bureau of Workers' Compensation (BWC) is developing a process for creating periodic outcome measures for providers who treat injured workers in Ohio.

The purposes of the outcome measures are: (1) to educate providers regarding their relative performance in treating work-related injuries so they can achieve better outcomes, (2) to comply with legislation establishing the Health Partnership Program, and (3) to develop an acceptable, credible procedure that may eventually be the basis for outcomes-based reimbursements and benefits to providers. As noted by the BWC "Procedure for Determining Physician Outcome Measurements for the Ohio Bureau of Workers' Compensation," the process must be valid, credible, and acceptable. The reviewers would add that it must also be useful in promoting system improvements.

As part of the development process, the BWC solicited an independent review of the proposed outcomes measures. The reviewers were provided with a summary of the procedure for determining physician outcome measurements, output describing system-wide outcomes for all providers, output showing individual provider analysis summaries for six providers, and a set of questions on the process developed by BWC (Appendices ...), as well as the original data that generated the outputs. At a later date, the reviewers were provided with a second set of provider analysis summaries, for the same six providers, in which cases defined as outliers had been eliminated

This report summarizes the comments and criticisms of the independent reviewers. Reflecting on the questions raised by BWC (Appendix A) and our own issues, this report addresses four major areas of concern, namely: validating the assumptions that generate the benchmark values; allocating costs to directing providers; adjusting for outliers and confounding variables; and producing analyses that are informative to both providers and administrators at BWC.

2. Validating Assumptions Generating the Benchmark Values

The BWC has established a system to assign an expected duration of work absence

(benchmark) based on the job classification of the injured worker and the most serious ICD-9 code allowed in the claim. The most serious ICD-9 code is defined as the code with the longest expected duration of work absence. The expected durations are based on the “Well Managed Benchmarks” from Milliman and Robertson Healthcare Management Guidelines, Volume 7. They have been adapted for Ohio by Milliman & Robertson by recalculating the expected durations for classification of occupation using NCCI codes instead of SOC codes. In addition, BWC has also developed expected durations for conditions not included in the Milliman and Robertson data using benchmarks from the Official Disability Guidelines.

The use of expected durations of work absence based on job classification is intended to adjust for any confounding of lost work time caused by any differences in the physical demands of the injured worker’s job. Assignment of a case to the most serious ICD-9 code is meant to adjust for injury severity.

This methodology raises a number of concerns:

- 1) the adequacy of using occupation as the single control for confounding
- 2) the credibility of the expected durations;
- 3) the accuracy of ICD-9 codes; and,
- 4) the use of the “most serious ICD-9 code” as indicator of severity.

It is true that differences in the physical demands of the employee’s work will significantly influence the duration of disability; patients with heavy jobs are likely to miss more time from work with a lumbar strain than patients with sedentary work. But there are a variety of other factors – outside the treating physician’s control – that have equally important influence on return to work. For instance, employer size can be important. A small employer may not be able to offer modified work or light duty to an employee with a heavy job, while a larger employer can. In this case, the two employees with lumbar strain will have very different durations of disability even though they both had heavy pre-injury jobs. Likewise a number of demographic factors - such as age, gender, education, and marital status - can have important effects on disability duration as well. There is no evidence presented that occupation is a better control for confounding than the other possible candidate variables. It may be that the use of occupation adds

complexity to the system without improving its accuracy.

The expected durations of disability are the “heart” of the system. But, no information has been provided detailing the methodology used by Milliman and Robertson (an accounting and actuarial consultant firm) in determining the expected durations derived from their Healthcare Management Guidelines. The acceptance of the proposed outcomes measurement methodology ultimately rests on the validity of these expected durations, and that, in turn, depends on the methodology and data used to generate them: Was it scientifically sound? Were the results tested for accuracy, reproducibility, and predictive utility? And was the development process transparent and open to critical review? No evidence addressing these issues is provided in the materials submitted for review.

Even if the expected durations are credible, their assignment to individual cases must be accurate. However, the ICD-9 codes found on provider billing statements and third party forms are almost never validated by an experienced nosologist before submission. Many of the most common conditions seen in the workers’ compensation system can be coded in multiple ways, entirely at the discretion of the health care provider. For example at least a dozen codes can be used for regional low back pain; some refer to symptoms, some refer to mechanisms of injury, and some to putative underlying pathophysiology. Individual physicians often have idiosyncratic predilection for one or two particular alternatives. In particular two physicians seeing the same patient may diagnose a lumbar strain and a lumbar disc herniation based on the same history and clinical findings, differing only in their personal beliefs about the “cause” of low back pain. And there are usually systematic differences between health care provider disciplines, especially between chiropractors and allopaths. There are no established criteria to guide health care providers in assigning ICD-9 codes and very little if any formal training is provided either in the professional schools or postgraduate training.

The severity of this problem is mitigated if either 1) ICD-9 codes are validated against medical records using explicit, standardized criteria; or 2) there are minimal differences in the expected durations associated with ICD-9 codes that are likely to be used

interchangeably. If ICD-9 codes are not validated and there are substantial differences between expected durations, then the use of the “most serious” ICD-9 code may result in significant misclassification of the claim. For example, a claimant being seen by the treating provider for hand pain consistently coded as “wrist sprain” may be seen on one occasion by an urgent care physician who codes the condition as “carpal tunnel syndrome.” The use of the most serious code could also make the outcomes measurement process susceptible to future “gaming”; providers wanting to improve their profiles could deliberately use the ICD-9 code with the longest expected duration of disability.

Absent independent validation of ICD-9 codes against medical records, this problem might be reduced by considering only the diagnostic codes submitted by the provider to whom the case is assigned and/or using more sophisticated protocols that consider all of the related ICD-9 codes found in the claims data and their relative frequencies. Alternatively, groups of related codes could be assigned a single expected duration of disability.

Finally, assuming that the expected durations are valid and the ICD-9 code assignment is accurate, there is still a concern that predictions are based on the most serious ICD-9 code, even in cases with multiple injuries (a legitimate reason for multiple ICD-9 codes). This method assumes that conditions do not interact with each other during the healing process; and it may fail to capture the potential impact of multiple injuries. Workers with multiple injuries may take longer to recover, on average, than is predicted from the single most serious injury. If so, the proposed process generates biased predictions of expected disability days for workers with multiple injuries.

In previous work (cites) one of us (MB) has used the ICD-9 code accounting for the largest proportion of medical payments as a proxy for severity. It would be informative to compare the two approaches. Does the ‘most serious’ ICD-9 code account for the majority of costs in most cases? If not, what injury groups/diagnostic codes are most likely to produce differences between the two severity proxies and what are the possible implications for the outcomes assessments?

3. Allocating Costs to a Directing Provider

Many injured workers receive care from more than one provider over the course of a workers' compensation claim. An important issue common to all attempts to construct provider assessments is how to identify the provider who has primary responsibility for the outcomes of a claim. The process proposed by BWC assigns a claim to a given provider if the provider has 60 percent or more of the Evaluation and Management (E&M) codes submitted for reimbursement over a specified time period (not stated in the materials provided), and holds the directing provider accountable for all costs and outcomes associated with the claim. Individual providers are analyzed if they have been identified as the directing provider in a minimum of 50 claims.

The use of E&M codes to identify treating health care providers is reasonable. E&M codes are used to bill office visits at which diagnoses are made and treatment planned. All provider types and specialties that can provide independent care to patients use E&M codes to bill for their office visits. Furthermore, many other services derive directly from office visits – laboratory tests and medical imaging studies are ordered, medications and therapies are prescribed – and are the responsibility of the treating provider billing the office visit. However, assigning the entire claim, with all of its costs and outcomes to the treating provider with the majority of the E&M codes is not reasonable.

First of all, 44% of claims remain unassigned because no one provider is responsible for more than 60% of the E&M codes. And then, there is no reason to believe that all of the claims assigned are actually assigned to the provider who was principally responsible for the majority of costs and the outcomes. For example, since surgical services are not classified as E&M, then a primary care physician who refers a claim to a surgeon after a period of conservative care could be identified as the directing provider even though the surgeon is responsible for the majority of costs and directs the claim after referral.

The 60 percent rule also appears to be an arbitrary break point and is not an appropriate indicator for chiropractors and some osteopaths. Chiropractors will conduct an evaluation at the onset of treatment, which will be billed as an E&M service. After that the

chiropractor will see the patient on multiple occasions to provide treatment, billing for manipulations and ancillary services but not an office visit. Patients treated initially by chiropractors may be incorrectly assigned to a later health care provider because of their relatively infrequent use of E&M codes in comparison to an MD, even though the chiropractor may have been the sole provider for months.

To enhance the credibility of its process, the BWC should provide more information regarding how the 60 percent cutoff was selected. Were other values tried and rejected? If so, why was 60 percent preferred? The BWC could perform sensitivity analyses using other values to determine how the conclusions/recommendations change with other break points.

A variety of strategies could also be employed to minimize these problems. Claims could be divided into different segments, based on the typical natural history of claims, and responsibility for each could be assigned separately. For instance, the initial period of conservative care can be separated from surgical care (if the claim proceeds to surgery), and surgical care from post-surgical rehabilitation, etc. E&M codes can then be used to determine the directing provider for each segment. Costs and outcomes for the segment could then be attributed to that provider. Alternatively, a fixed period of observation, for example the first 90 days of treatment, can be used and the directing provider determined and assigned responsibility for costs and outcomes of that part of the claim. Or a “moving window” of observation of 30-60 days could be employed, with a directing provider assigned responsibility for the costs and outcomes as long as he/she billed >80% of the E&M codes; that provider’s responsibility would end and attribution of further costs and outcomes would shift to another provider if he/she became the source of >80% of the E&M codes.

The 56% of claims that are assigned a directing provider by the BWC’s proposed methodology is further reduced to 45% of the original sample by the application of the requirement that a provider have at least 50 claims to be included in the final analysis. The reduction in the number of providers is even more dramatic. An important concern

with respect to the credibility of the proposed process is that only 2.5 percent of active providers will be analyzed under current rules. One possible approach to include more providers is to relax some restrictions, but distinguish between confidence levels in the reports. For example, define Level One confidence as 60% E&M codes, with 50+ cases; Level Two confidence as 60% E&M codes, 30+ cases; Level Three confidence as 50-59% E&M codes, 50+ cases; and so forth.

Another concern in defending the credibility of the proposed process is the small fraction of total medical costs paid to directing providers. On average, directing providers receive only 28 percent of total medical payments for medical only claims, and only 14 percent for lost time claims (Sheet II, Chart B). The percent of medical payments paid to the directing provider also varies widely across provider types. On average, directing providers who are medical doctors receive 23 percent of medical payments, doctors of osteopathy receive 22 percent, podiatrists receive 29 percent, and chiropractors receive 68 percent (Sheet II, Chart D).

The reporting format in the examples of provider analysis summaries provided the reviewers implies that the directing provider is responsible for the other two medical cost outcomes. This may or may not be reasonable. Certainly a provider is ultimately responsible for the costs of medication and therapies prescribed as well as diagnostic tests and medical imaging studies ordered. The directing provider could also reasonably be assigned responsibility for the office visit costs of consultant referrals he/she instigated. But Ohio is an employee choice state; employees can self-refer for evaluation and care. When an employee's care shows evidence of multiple providers, it becomes unclear who is responsible for which costs. Even when the primary care provider has made a surgical referral it strains credibility to make that provider responsible for the surgery performed by the consultant and the post-operative treatment costs. The BWC should explore if there is any way to estimate the actual percent of costs reasonably attributable to the directing provider.

4. Adjusting for Outliers

In developing the provider analysis summaries, the BWC proposes to eliminate outliers from each provider's claim distribution so that a provider is not penalized for factors beyond his/her control. Outliers are defined by comparing actual to predicted durations of work absence based on the most serious ICD-9 code. Three criteria for identifying outliers have been proposed: (1) claims for which the difference between actual and expected duration of absence is more than two standard deviations above the mean difference for the provider; (2) the 15 percent of worst cases for each provider; (3) a standard number of worst cases for each provider. Output submitted to the reviewers defines outliers according to the first criterion.

First of all, this proposed correction assumes that the only reasons for outliers are factors beyond the health care providers' control. In fact, outliers could just as easily be indicators of poor treatment. Any corrections should ultimately allow the system to retain information on how many outliers occur in a provider's data and how often. This can then be compared to expectations based on the experience in the population as a whole, thus allowing identification of providers with unexpectedly high numbers of outliers.

Next, the alternate criteria for excluding outliers have quite different potential impacts on the provider assessments. Excluding cases with differences between actual and expected durations two or more standard deviations above the mean difference eliminates an unpredictable and arbitrary percent of cases for each provider. In the output submitted to the reviewers, this varies from 1.5 (JK) to 9 (PS) percent of cases. Eliminating a fixed (15) percent of cases for each provider penalizes providers who treat more severe cases (differences in severity that are not adequately reflected in the assignment of expected duration by most serious ICD-9 code and occupation), whereas eliminating a standard number of cases places large-volume providers at a disadvantage.

The proposed exclusion rules are also an unusual means to control for confounding because the rules are based on outcomes (dependent variables) rather than inputs (independent variables). The exclusion rules may, therefore, be quite inadequate to adjust for differences in the case mix treated by different providers. This appears to be the case

in the preliminary output sent to reviewers (which uses criterion #1 to exclude outliers). The preliminary data includes provider analyses for one chiropractor and four medical doctors, and shows that the chiropractor's case mix differs dramatically from the case mix treated by medical doctors (e.g. no loss time claims, no outliers, no claims with durations above the optimal). The results suggest that provider comparisons should be restricted by provider type, but the claim count for the chiropractor is extremely small (N=8), so this issue should be reconsidered with additional summaries for different providers.

Even restricting comparisons within provider types may be inadequate to control for differences in case mix and other confounding variables. Comparisons of the four MDs in the preliminary data suggest that one physician (PS) has a distinctly different case mix than the others. PS treated 80 percent loss time claims, with an average cost of \$28,400, compared to 5-12 percent loss time claims for the other three physicians, with average costs less than \$1500. The distribution of claim types does not account for the higher average costs for PS, as her average costs for MO claims are more than triple the average costs of MO claims for the other physicians, and her average costs for loss time claims are more than quadruple their average costs. According to the outcomes measures PS performs poorly: 70 percent of cases beyond goal (7-14% for others), with the mean time beyond goal in excess of 15 months (2-5 months for others). Eliminating the nine percent outliers for PS improves her outcomes measures, but does not begin to bring her performance in line with other physicians. Given such striking differences, I would want to know much more about practice characteristics for PS relative to other physicians in the sample before concluding that such comparisons provided a credible review of her performance. Restricting the analyses to claims for the same conditions, e.g. back sprains or strains, carpal tunnel syndrome, may produce more credible provider comparisons.

Given that an acceptable mechanism for identifying outliers is defined, both adjusted (typical claims) and unadjusted (all claims) outcomes should be reported. The unadjusted measure is more appropriate for improving system outcomes, because outlying claims account for a disproportionately large proportion of overall costs. The adjusted measure

is more valid for comparisons across physicians because case mix is standardized to some degree.

5. Producing Valid and Informative Provider Analyses

Comparing an individual physician's outcomes (health care and disability costs, durations of work absence) to means for his/her peers may not be a good indicator of satisfactory performance. This is because distributions of health care costs and work absences typically are skewed to the right, and the mean is sensitive to the tails of the distribution. Note that the distributions of total days absent for all provider groups (Sheet I, Chart C) follow this skewed pattern. For example, 77% of claims treated by an MD have durations of absence within the optimal predicted value, and the average number of days absent for these claims is within 6 days of the optimal. Only 17% of claims exceed the optimal, but the average number of excess days is 145. These 17% of claims account for a disproportionate share of health care and disability costs, skewing the cost distributions as well. The outlying values have a strong impact on the mean, so individual physicians who have costs at or slightly below the mean are not performing well relative to the great majority of claims. A better approach may be to compare a physician's outcomes to the median of the cost distributions for his/her peers, because the median is much less sensitive to outlying values.

Appendix B shows a report format with a number of ratios that would be useful for provider analyses, which are not directly reported on the provider summaries. These include: percent outliers identified in the all claims data, distribution of claim types (percent of loss time claims), percent of cases in which the injured worker has returned to work, and percent of cases with durations of work absence beyond the benchmark. It is unclear to what extent these variables are outcomes measures, and to what extent they describe differences in case mix and other confounding variables.

The proposed "Provider Analysis Summary" is both overly complex and too vague. First, the results are divided into too many categories of questionable relevance. Results are reported for claims with insufficient data; almost by definition, there is no value to

outcomes measurements derived from cases with insufficient data. Presentation of results based on insufficient data is at best a distraction and at worst could invite dismissal of the exercise. Moreover, claims are divided into four types, two of which – RL and RM – have too few instances to be of use in comparison across providers; these types of claims should be excluded from the presentation. And, a distinction is made between cases with RTW and without. The RTW indicates that the patient is back at work and disability is ended. Lack of an RTW implies that lost workdays are still accruing. This distinction while important to the system’s administrators has no clinical relevance and will be unimportant to health care providers if not misunderstood. Furthermore, lost workdays are reported for both “medical-only” claims and “lost work time” claim. Since the outcome measure is lost days, whether or not they were compensated, the distinction between MO and LT claims is irrelevant to the objectives of the report and simply adds complexity to the presentation. The essential information is the number and proportion of claims that are within goal and beyond goal; this should be highlighted by the presentation of the data.

Next, medical cost outcomes are reported without any frame of reference or benchmark so that readers are unable to judge the provider’s performance. Medical costs are also divided into those paid to the provider, drug costs, and payments to others. Without benchmarks these categories are no more informative than total medical costs. Construction of benchmarks will require differentiation by provider type, and within MDs by specialty, since the costs paid to the directly to the provider will differ significantly (predominantly E&M services for primary care providers, predominantly physical medicine services for chiropractors, while surgery charges will account for most costs paid to surgeons). As noted above there are significant methodological concerns regarding which costs can legitimately be attributed to the directing provider.

Finally, aggregate results such as those on the “Provider Analysis Summary” examples may be too vague to have more than very limited utility as feedback to practicing physicians on their practice. A physician will find out that 20% of their claims lose more time than expected. But, how do they compare to other providers, especially of the same

discipline and specialty. And do they recommend too much treatment and time off for everyone or only for patients with certain conditions? A more useful approach would report outcomes at least for those conditions that account for most of the problem outcomes in the entire population of claims (e.g. likely to include low back pain claims). This will help direct the healthcare provider to areas of their practice that can be improved.

Template for Provider Analysis Summaries

Appendix C presents a draft template for comparing provider summaries to benchmarks based on results for all providers. As noted above, the reviewers suggest that the most credible benchmarks will be based on comparisons within provider type and injury category. The draft template is restricted to medical doctors.

The template is restricted to a limited number of fields, relative to the data provided for individual providers (Appendix B), because much of the relevant information was not included in the provider summary data. Most importantly, the provider summary data obscures differences between MO/LT claims that may explain some of the variations across providers. The following data fields are needed in the provider summary reports to produce a more complete analysis: number of providers within each provider type, mean days beyond goal reported separately for MO/LT claims, mean days absent separately for MO/LT claims, mean costs separately for MO/LT claims, and percent outliers in the overall data. Percent outliers across all providers can be defined to be consistent with the individual provider reports: claims for which the difference between actual and expected duration of absence is more than two standard deviations above the mean difference for all providers.

The proposed template compares results for individual MDs with overall means for the provider category. We also recommend giving providers information to evaluate where their outcomes fall in the overall distribution of their peers. A first step is to report standard deviations as well as means in the all-provider results. To make the comparisons more user-friendly, BWC can report ranges for each mean encompassing values from -1

to +1 standard deviations from the mean, and -2 to $+2$ standard deviations. If the distributions are approximately normal (and this should be tested) 68 percent of providers will have outcomes in the first range, and 95 percent in the second.

Appendix A

Questions from Ohio BWC

- 1. Does the use of dates of service for E & M Codes appear reasonable as an identifier that the physician is managing the claim? Is so, why? If not, what recommendation would you offer?*
- 2. Does the 60% or greater E & M Codes seem reasonable to identify claims assigned to a given provider? If so, why? If not, what recommendation would you offer? Would you recommend another value?*
- 3. While realizing that case severity is complex and difficult to address and realizing that management of disability (lost time) is one of the key outcomes, does the use of the most serious ICD-9 Code combined with the NCCI or SOC code seem to represent a reasonable estimate of the severity of the case that could be determined and utilized from available data elements? If so, why? If not, what other factors should be considered?*
- 4. The current process plans to eliminate the worst cases for each provider to adjust for confounding. Currently we plan to remove 15%. Assuming this process is to be used, is the 50 case population size proposed an adequate sample size to obtain a credible representation of the physician's outcomes? If not, what population size would you recommend?*
- 5. The current process is to eliminate the worse cases for the provider so the provider is not penalized for factors beyond his control. Should the cases eliminated be the worst 15% of cases as proposed, those cases defined as exceeding the duration of disability by more than two standard deviations, a cap of "X" number of cases, or another method? Please provide any recommendations and comments.*
- 6. The current proposal has an actual outcome as measured by the number of lost versus expected lost work days without removal of the worse 15% and an "adjusted outcome" which is the outcome with removal of the worse 15%. Assuming that some of the worse cases are removed from the provider pending the recommendations of Question 5, what are your opinions of reporting both an actual and an "adjusted" outcome and which is more valid?*
- 7. Expected or benchmark values are only available for the expected lost number of work days based on the most severe ICD-9 Code allowed in the claim coupled with the NCCI or SOC job classification. From the population of claims for each provider, the proposal is to provide the total and average claim costs for temporary total payments, medical payments to the provider, pharmacy costs, and total medical payments. Are there other outcomes that should be reported?*

8. *Please comment on the credibility and validity of this process in terms of representing an outcome of the medical management of the claim by a physician and to be used as an education tool to try to generate improvement.*
9. *Please describe any limitations or concerns of the process for use in certification of physicians or for use in an incentive or bonus payment system.*

Appendix B
Compilation of Provider Analysis Summaries – MD Only

All claims

<i>Provider</i>	<i>%outliers</i>	<i>%LT</i>	<i>Average costs</i>			<i>%beyond goal</i>	<i>%return to work</i>	<i>Mean days absent</i>		<i>Mean days beyond goal</i>		<i>N</i>
			<i>All</i>	<i>MO</i>	<i>LT</i>			<i>MO</i>	<i>LT</i>	<i>MO</i>	<i>LT</i>	
PS	9%	80%	\$28,400	70%	\$34,915	70%	66%	1.6	442	4	463	181
PM	4%	5%	\$877	7%	\$5550	7%	96%	0.5	150	4	168	446
EI	3%	12%	\$1489	14%	\$7396	14%	95%	1.0	58	3	61	575
JK	1.5%	7%	\$912	7%	\$5021	7%	98%	0.6	60	11	69	1607

Source: OBWC Data Warehouse, Provider Analysis Summaries.

Typical claims

<i>Provider</i>	<i>%LT</i>	<i>Average costs</i>			<i>%beyond goal</i>	<i>% return to work</i>	<i>Mean days absent</i>		<i>Mean days beyond goal</i>		<i>N</i>
		<i>All</i>	<i>MO</i>	<i>LT</i>			<i>MO</i>	<i>LT</i>	<i>MO</i>	<i>LT</i>	
PS	78%	\$25,273	69%	\$31,734	69%	73%	1.6	339	4	365	181
PM	4%	\$722	6%	\$3488	6%	97%	0.5	33	4	27	446
EI	10%	\$1025	12%	\$4085	12%	96%	0.9	25	3	22	575
JK	6%	\$835	6%	\$4057	6%	98%	0.5	25	8	23	1607

Source: OBWC Data Warehouse, Provider Analysis Summaries.

Appendix C
Provider Outcomes Compared to Means for All Providers – MD

	N	% lost time	% beyond goal	% return to work	Mean cost
All MD		17%	17%	93%	\$2544
PS	181	80%	70%	66%	\$28,400
PM	446	5%	7%	96%	\$877
EI	575	12%	14%	95%	\$1489
JK	1607	7%	7%	98%	\$912
<i>Source: OBWC Data Warehouse, Provider Analysis Summaries and Totals/All Providers.</i>					