The Journal of Mental Health Policy and Economics
J Ment Health Policy Econ **5**, 141-152 (2002)

Consistency in Performance Evaluation Reports and Medical Records

Mingshan Lu[1]* and Ching-to Albert Ma[2]

[1]*Assistant Professor, Department of Economics, University of Calgary, Canada*
[2]*Professor, Department of Economics, Boston University and Hong Kong University of Science and Technology*

Abstract

Background: In the health care market managed care has become the latest innovation for the delivery of services. For efficient implementation, the managed care organization relies on accurate information. So clinicians are often asked to report on patients before referrals are approved, treatments authorized, or insurance claims processed. What are clinicians' responses to solicitation for information by managed care organizations? The existing health literature has already pointed out the importance of provider "gaming," "sincere reporting," "nudging," and "dodging the rules."
Aims of the Study: We assess the consistency of clinicians' reports on clients across administrative data and clinical records.
Methods: For about 1,000 alcohol abuse treatment episodes, we compare clinicians' reports across two data sets. The first one, the Maine Addiction Treatment System (MATS), was an administrative data set; the state government used it for program performance monitoring and evaluation. The second was a set of medical record abstracts, taken directly from the clinical records of treatment episodes. A clinician's reporting practice exhibits an inconsistency if the information reported in MATS differs from the information reported in the medical record in a statistically significant way. We look for evidence of inconsistencies in five categories: admission alcohol use frequency, discharge alcohol use frequency, termination status, admission employment status, and discharge employment status. Chi-square tests, Kappa statistics, and sensitivity and specificity tests are used for hypothesis testing. Multiple imputation methods are employed to address the problem of missing values in the record abstract data set.
Results: For admission and discharge alcohol use frequency measures, we find, respectively, strong and supporting evidence for inconsistencies. We find equally strong evidence for consistency in reports of admission and discharge employment status, and mixed evidence on report consistency on termination status. Patterns of inconsistency may be due to both altruistic and self-interest motives.
Discussion and Limitations: Payment contracts based on performance may be subject to provider mis-reporting, which could seriously undermine its purpose. However, further analysis is needed to determine how much of the inconsistencies observed are results of clinician gaming in reporting.
Implications for Health Policy: Increasing system accountability is becoming more and more important for health care policy makers. Results of this study will lead to a better understanding of physician reporting behavior.
Implications for Future Research: Our work in this paper on the data sets confirms the statistical significance of strategic reporting in alcohol addiction treatment. It will be of interest to confirm our finding in other data sets. Our on-going research will model the motives behind strategic reporting. We will hypothesize that both altruistic and financial incentives are present. Our empirical identification strategy will use Maine's Performance-Based Contracting system and client insurance sources to test how these incentives affect the direction of clinician's strategic reporting.

Received 27 November 2002; accepted 9 April 2003

Introduction

In the health care market managed care has become the latest innovation to ensure efficient delivery of services. Managed care organizations often regard quantity restrictions as the key for controlling moral hazard. Here, it is important that quantity control does not become too excessive. For efficient implementation, a managed care organization relies on accurate information. So clinicians are often asked to report on patients before referrals are approved, treatments authorized, or insurance claims processed.

What are clinicians' responses to solicitation for information by managed care organizations, regulators, or insurers? Do clinicians always provide consistent information on clients? Do clinicians' reports on a client's treatment episode differ depending on the intended use of the information? If their reports are inconsistent, how is the discrepancy characterized? On which aspects of a client's treatment will consistency in clinician reports be more likely? It is important to address these issues because the successful implementation of managed care relies on the quality of information supplied by clinicians.

The existing health economics literature has already pointed out the importance of provider "gaming," "sincere reporting,"

***Correspondence to:** Prof. Mingshan Lu, Department of Economics, University of Calgary, 2500 University Drive, NW, Calgary, AB, Canada, T2N 1N4.
Tel.: +1-403-220 5488
Fax: +1-403-282 5262
E-mail: lu@ucalgary.ca

Source of Funding: Research was funded by grant R21-AA12886 from the National Institute on Alcohol Abuse and Alcoholism; Lu also gratefully acknowledges financial support from the Alberta Heritage Foundation for Medical Research and the Institute of Health Economics.

"nudging," and "dodging the rules".[1-3] * In this paper, we present statistical evidence for report inconsistency. Our empirical work leads to the fundamental question of the motive behind report inconsistencies. We can identify two such motives: altruistic and self-interest.

First, a clinician may report strategically because of his altruistic motive. For example, feeling that managed care controls interfere with care, clinicians may exaggerate the severity of an existing problem at admission. This may make for an easier approval for request. As another example, a clinician may report an addiction severity lower than the true level at the time of discharge, helping the client to avoid or reduce the negative social stigma associated with alcohol abuse. Second, a clinician may report strategically because of his self interest. The clinician's reward, either monetary or professional, may be based on reported outcome measures.[6] Financial or professional success or advancement may come with good clinical performances. A clinician may work hard for such good outcomes, but another way to "improve" a performance is simply to report a better one.[7] Misreporting information for personal gain is a common strategy in many social situations.

The current paper aims to look at only the evidence for report consistencies; we defer the theoretical issue of the altruistic and self-interest motives and their empirical identification to a continuing research effort. In this paper, we present two unique data sets for assessing the consistency of clinicians' reports. We can investigate directly information consistency because the two data sets contain reports made by the same clinician on the same treatment episode for two different purposes. These two data sets are on approximately 1,000 alcohol treatment episodes for the period 1990-1995 in the state of Maine. The first one, the Maine Addiction Treatment System (MATS) data, was an administrative data set; the Maine Office of Substance Abuse (OSA) used it for program performance monitoring and evaluation. The second was a set of medical record abstracts, taken directly from the clinical records of treatment episodes.

We look for evidence of systematic inconsistencies in five categories: admission alcohol use frequency, discharge alcohol use frequency, termination status, admission employment status, and discharge employment status. All five measures are used by the state of Maine to monitor treatment performance. For the first two measures, admission and discharge alcohol use frequencies, we find, respectively, strong and supporting evidence for inconsistencies. On the other hand, we find strong evidence for the lack of inconsistencies in the last two. There is also mixed evidence on report consistency on termination status. Patterns of inconsistency may be due to both altruistic and self-interest motives.

Our decision to investigate these five measures reflects our initial beliefs on where consistency or inconsistency likely occurs. As indicators of health and social functioning status or treatment outcomes, all five measures are potentially subject to misreporting. However, misreporting about a client's employment status can be problematic because employment information is readily verifiable; a physician risks embarrassment or financial penalties which may result from government potential audits, or colleagues who happen to find out. Alcohol use frequencies are more personal information; most clients are unwilling to disclose the information except to medical personnel. A clinician may be more willing to misrepresent this "privileged" information. Treatment termination status is less straightforward information than employment status, but perhaps more easy to verify than alcohol use frequencies.

Our use of two data sets is a significant improvement over previous research on similar investigations. If access to two reports made by a clinician is infeasible, then the question of report consistency can only be investigated indirectly. Often researchers use an independent clinician to review patient information reported by a clinician to determine if there has been any bias. To control for the variation in practice and reporting styles among clinicians, some sort of differencing needs to be used (see Carter *et al.*[8] for an illustration). Our direct method is straightforward, and does not involve judgment by an outside reviewer. The differencing control in the indirect method is imperfect, and our method avoids that.*

In this paper, we do not need to assume that one of the two data sets is more reliable. A priori, one would expect that medical records could be more trustworthy, whereas the clinician's administrative report to OSA might be more prone to manipulation. Withholding or omitting information in medical records is professionally and ethically unacceptable, may have legal consequences, and adversely affects care should the patient be transferred to other clinicians. Moreover, we do not believe that clinicians in our sample anticipated that their medical records would be examined in a way that might conflict with their own interests. By contrast, a clinician's report to OSA was for evaluation purposes. Furthermore, in 1992, OSA implemented Performance Based Contracting (PBC), under which providers were compensated according to their clients' treatment outcomes as reported in MATS. Our concern, however, is to document the inconsistencies in two reports made by the same clinician on the same episode within the sample period. For our purpose, it is not necessary to ascertain that medical records are a gold standard.

Data

We use two data sets: the Maine Addiction Treatment System (MATS) data and a medical record abstract data collected by Boston University researchers in 1996. MATS contains information on all clients served by substance abuse treatment programs that received funding from the federal government or the state of Maine. Covering the period between October 1,

* Novack et al[4] reported that 87% of respondents of a survey on physicians regarded deception as an acceptable way to help patients on occasion. Carter *et al.*[5] estimated that about 30% of the increase in Medicare expenditure between 1986 and 1987 was due to case mix index upcoding.

* In breast cancer research, researchers have compared administrative data, such as Medicare claims files and hospital discharge data, with medical records and patient surveys.[9-18] These studies focused on matching patients in administrative data with those in medical records and vise versa.They tested whether information in administrative data could be used to measure pattern of care such as type of surgery and length of stay.

1989 and June 30, 1998, MATS data are submitted by clinics on standardized admission and discharge forms, and reported to Maine Office of Substance Abuse (OSA) for treatment performance evaluations.

MATS is collected in the following way. A record in MATS is based on a treatment episode defined by a clinician. When a client is admitted (or readmitted) to a treatment program, he or she will be interviewed and asked a series of standardized questions. The client's answers are recorded in a MATS admission form and any required program-specific information is added. When the client completes treatment and leaves the program, he or she will be interviewed again during the last visit. Answers of the clients are recorded in a standardized MATS discharge form and this last visit defines the end of a treatment episode. When observing that a client has not come for treatment for a long period of time, a clinician will obtain information from the clinical records of this client's last contact and fill a MATS discharge form. MATS did not impose a uniform standard on when the admission and discharge forms must be filled, but did require that "the counselor having the face-to-face contact with the client" complete the forms "either during the session or soon after".[19] Nevertheless, it is our understanding that administrative staff at some programs might have completed the MATS forms based on information collected in interviews or in clinical records.

In both interviews, MATS collects information about demographics (age, race, sex, education, living situations), income, employment status, criminal involvement, and health variables (pregnancy, recent medical treatments, alcohol use), as well as a client's substance abuse severity such as types of alcohol, frequencies, routes of administration, and ages of first use of primary, secondary and tertiary substances. Each treatment episode in MATS notes the type of treatment program and provider. Service delivery information, such as the number of treatment units and unit cost, is recorded. Finally, MATS identifies the termination status of a client, defined as the main reason of terminating the treatment. The identified reasons are: completion of the treatment, referred (further treatment is not appropriate for client at this facility), client discharged without clinic agreement (i.e. client leaves without explanation), noncompliance with rules and regulations and/or client refuses service/treatment, deceased, incarcerated, moved out of a catchment area, or discharged due to program cut/reduction.

The medical record abstract data was collected by Boston University researchers under the supervision of OSA representatives in the summer of 1996. A project manager of a grant supported by the National Institute of Alcohol Abuse (NIDA) and two research assistants were responsible for the abstraction of data from actual medical records. The project manager was experienced with the MATS data; she was the manager for the entire five-year duration of the NIDA grant. The two research assistants were a graduate student and an advanced undergraduate student. They were trained by the project manager before the actual data collection; samples of the MATS form and clinical records were given to them for review. At the first few sites, some double coding was implemented to ensure that the same standard was used even though no formal inter-rater reliability test was conducted. A uniform standard was adopted for the entire collection process.

The record abstract data set consists of information on 988 treatment episodes covering the period from October 1990 to June 1995. These episodes are randomly selected from MATS according to the following criteria on the clients and the providers. First, clients must have alcohol abuse at admission as a primary problem, and must receive outpatient treatments. Second, to avoid potential selection problems, clients must have had no prior treatment experience one year before the beginning of their current treatment. Third, the episodes come from programs at ten large providers (clinics). The actual data points are obtained by sampling a hundred episodes evenly distributed across each fiscal year. These episodes satisfy the requirements on the client side and are from the ten large providers. Out of these 1000 episodes collected, 12 were excluded due to incorrect client identification, missing clinical record, or duplicated abstraction. This resulted in a total of 988 episodes in the record abstract data.

After researchers identified those clients and episodes for the record abstract data set, OSA was notified and asked to provide supervised access to the corresponding clinical records. Officials at OSA then used a scrambling algorithm to identify the requested clinical records. In the summer of 1996, Boston University researchers went to clinics where the clinical records resided. Identified records were then reviewed and information of a list of variables was noted down by researchers.

Clinical records are hand-written, free-format document written during or after each patient's visit. The variables in the record abstract data set include admission and discharge dates, the number of taken visits, their exact dates, whether appointments have been kept, the title of the responsible clinician, and the type of treatment in each visit. The health status measures in the record abstract data set are frequency of use at admission and discharge, whether abstinence is a stated goal, relapse after the previous visit, reduction of use, and whether abstinence is achieved at discharge. The record abstract data set also contains clinicians' private judgment on clients' progress towards abstinence or other treatment goals in each visit. The client's termination status as well as employment status at time of admission and discharge are recorded.

We must note that our two data sets were intended for different purposes. The MATS form was a structured design, while the medical records were free-format, hand-written documents. Nevertheless, these two data sets were the only sources of information for studying information consistency in our sample. The data collection methodology was to take extra care in extracting information from medical records to minimize potential problem due to the difference in data characteristics. This has led to a decision to regard some information as missing when there was any doubt about that information in the medical records. We have chosen to use multiple imputation as a sensitivity check; more discussions can be found in Sensitivity Analyses section. In any case, the

Table 1. Characteristics of Clients

Client's characteristics	Percent (n=988)	Mean	S.D.
Age		31.76	11.63
Male	73.91		
Marital Status			
Married	21.32		
Divorced/Widowed/Separated	32.18		
Single/never married	46.50		
Education (years)		12.37	2.22
Employment			
Full time	28.63		
Not Full time	71.27		
Legal Status			
With legal involvement at time of admission	53.50		
Concurrent Psychiatric Problem	12.59		
Household income (last 30 days)		$ 856.21	$ 847.60
Primary Payer Status			
OSA	26.90		
Medicaid	22.84		
Self-pay	23.65		
Privately-insured	18.78		
Other	7.82		
Admitted after PBC is implemented	63.53		
Discharged after PBC is implemented	70.36		
Alcohol Use Frequency at Admission			
Moderate user	60.20		
Heavy user	39.80		
Severity of Alcohol Abuse			
Casual/Experimental user	5.89		
Lifestyle-involved user	21.73		
Lifestyle-dependent user	38.78		
Dysfunctional user	19.70		
Undetermined	13.91		
Number of Prior Treatment Episodes			
No prior treatment episodes	50.25		
One prior treatment episodes	27.92		
Two or more prior treatment episodes	21.83		
Termination Status			
Completed treatment	35.73		
Referred	8.60		
Without clinic agreement	36.34		
Died	0.20		
Incarcerated	0.71		
Moved/can't attend	4.96		
Noncompliance/refused treatment	11.23		
Discharged due to program cut/reduction	0.51		
Unknown reason	1.72		

Notes:
1. Information reported in MATS is used. Percentages are reported for binary variables; means and standard errors are reported for continuous variables;
2. Moderate users include those who drink once per month, two to three days per month, once per week, or two to three days per week; heavy drinkers include those who drink four to six days per week, once per day, two to three times per day, or more than three times per day.[20]

record abstract data set is merged with the MATS data. The merge allows us to crosscheck information on the same client from the administrative MATS and clinical record abstract data sets. In sum, for each episode, we have information from two different data sets.

Data Analyses and Results

Descriptive Statistics

The main characteristics of the 988 clients in our sample, as reported in MATS, are listed in **Table 1**. These clients were predominantly male and unmarried, with an average age of thirty-two and an average of twelve years of education. At the time of admission, less than one third of the sample were employed full time. More than half of the clients had legal involvement, and more than 10 percent had concurrent psychiatric problem. Average household income in the past 30 days before admission was lower than nine hundred dollars.

Less than twenty percent of the clients had private insurance. Medicaid, OSA, and clients' own resource each supports roughly 25 percent of all clients. Our understanding is that many clients who reported to pay treatment with their own resource (classified as "self-pay") would rely partly on state support. Up till early 1990's, Maine's government allocations to providers were based on historical funding levels, with yearly changes being spread evenly across providers according to changes in state and federal appropriations. On July 1, 1992, Performance-based Contracting (PBC) was implemented to allocate state funding based on provider performance.[6] Provider performance is measured using performance indicators defined by the state. These indicators are constructed using the information in MATS, including information on alcohol use frequency, termination status, employment status, etc.[21,22] More than sixty percent of our sample were admitted after PBC was introduced, about two third discharged after PBC was introduced.

Three measures of a client's alcohol abuse problem at admission are recorded in MATS. The first is a categorical measure of the client's alcohol use frequency at admission, coded in nine categories: not drinking in the past 30 days, drinking once per month, two to three days per month, once per week, two to three days per week, four to six days per week, once daily, two to three times daily, or more than three times daily. About sixty percent of our sample are reported as moderate users in MATS, defined as those drinking at least once per month but less than four days per week; the rest are heavy users who drink at least four days per week.* The second is a counselor-assessed alcohol abuse severity measure. About five percent of the sample are assessed as casual or experimental users, one fifth as lifestyle-involved user, two fifths as lifestyle-dependent users, and one fifth as dysfunctional users. The third measure is a client's number of prior treatments. About half of the sample had no prior treatment episode, about a quarter had one prior treatment episode, and the rest had two or more.

A course of treatment may be terminated for various reasons. Only about one third of clients in our sample completed treatment. More than one third left a treatment program without any explanation ("without clinic agreement"); another eleven percent refused treatment. The rest terminated treatment because they were referred, deceased, incarcerated, moved out of a catchment area, or discharged due to program cut/reduction.

Comparison between MATS and Record Abstract Measures

To examine clinician's reporting behavior, we compare five measures recorded in both MATS and record abstract data: *admission alcohol use frequency, discharge alcohol use frequency, termination status, admission employment status, and discharge employment status.* They reflect three dimensions of substance abuse treatment goals and effectiveness: to reduce a client's substance usage, to retain a client in the treatment program and ensure the treatment is completed, and to improve the client's social functioning.

All five measures are used by OSA to construct treatment effectiveness indicators. Compared with employment status, a client's alcohol use frequency is subject to a clinician's assessment and a clinician's private information; the likelihood of a misreport being discovered is much lower. Termination status is less straightforward than employment status, but certainly easier to verify than alcohol use frequency. In other words, if incentives exist for clinicians to manipulate reports in MATS, we expect that the clinicians will more likely manipulate reports on alcohol use frequency than on employment and termination status. We test the hypothesis that the degree of inconsistencies between the MATS and record abstract data is highest on the alcohol use frequency measures, and lowest on the employment status measures.

Table 2 presents the joint distribution of the admission alcohol use frequencies reported in the MATS and record abstract data. Each number in a cell reports the number of clients who have been classified in the corresponding alcohol use frequency categories in the MATS and record abstract data sets. As an illustration, consider the number 12 in the third row and the first column of **Table 2**. Of the 988 clients in our sample, 12 of them are reported by the record abstract data not to have a drink in the past 30 days and by the MATS data to drink two to three days per month. For these clients, the record abstract data and MATS data yield inconsistent information. Two cells to the right is the number 24. Here, these 24 clients are classified by both MATS and record abstract data to drink 2-3 days per month; the two data sets yield consistent information.

The last row is the marginal distribution of use frequencies according to the record abstract data; last column, according to MATS. Again, as an illustration, consider the number 67 on the last row. The record abstract data report a total of 67 clients who drink once a week. For these 67 clients, their

* We follow the definitions of moderate and heavy drinkers as those used in Lu and McGuire.[20]

Table 2. Cross frequency table of admission alcohol use frequency (MATS vs. Record Abstract)

		Record Abstract Data										
		none in past 30 days	once per month	2-3 days per month	once per week	2-3 days per week	4-6 days per week	once daily	2-3 times daily	>3 times daily	Missing	Total number
MATS	None in past 30 days	**0**	0	0	0	0	0	0	0	0	0	0
	Once per month	21	**21**	23	9	3	1	1	3	2	51	135
	2-3 days per month	12	10	**24**	13	7	3	2	0	2	45	118
	Once per week	8	4	10	**20**	21	7	1	1	4	29	105
	2-3 days per week	30	8	16	13	**58**	18	10	3	5	74	235
	4-6 days per week	8	6	5	4	19	**29**	7	5	9	33	125
	Once daily	13	0	7	4	5	11	**31**	7	8	32	118
	2-3 times daily	8	0	0	2	1	1	7	**18**	7	8	52
	>3 times daily	7	2	5	2	2	5	9	14	**34**	17	97
	Missing	0	0	0	0	1	0	0	0	0	2	3
	Total number	107	51	90	67	117	75	68	51	71	291	988

admission use frequencies according to the MATS data are in the same column. If the two data sets were completely consistent, all of these 67 clients would have been reported as drinking once a week by the MATS data. In fact, only 20 among these 67 clients were so reported by MATS.

The entries on the diagonal (in bold) of the table indicate the occurrences of consistent reporting on admission alcohol use frequencies between MATS and the record abstract data. Below the diagonal line, MATS reports higher admission alcohol use frequencies than the record abstract data. Conversely, above the diagonal line, MATS reports lower alcohol use frequencies. **Table 2** shows that both exaggeration and understatement of admission alcohol use frequency in MATS are common. Such a pattern is consistent with our hypothesis that strategic reporting is driven by both altruistic and self-interest motives.

The large percentage of missing values (29.45 percent) in the record abstract data is a result of a cautious methodology in data collection. Our research team had to go through actual medical records to collect the information. Being hand-written notes in free formats, medical records could be difficult to understand. Where the uncertainty about accuracy of information was judged to be significant, our research team members reported the information as unavailable. In the next

Table 3. Cross frequency table of discharge alcohol use frequency (MATS vs. Record Abstract)

		Record Abstract Data										
		none in past 30 days	once per month	2-3 days per month	once per week	2-3 days per week	4-6 days per week	once daily	2-3 times daily	>3 times daily	Missing	Total number
MATS	None in past 30 days	**465**	7	7	6	5	2	0	0	0	91	583
	Once per month	10	**12**	6	3	0	0	1	0	0	30	62
	2-3 days per month	11	2	**6**	4	2	2	0	0	1	30	58
	Once per week	8	0	1	**7**	10	1	3	0	2	35	67
	2-3 days per week	10	3	3	3	**9**	2	3	2	0	47	82
	4-6 days per week	2	0	2	3	5	**9**	1	0	0	25	47
	Once daily	6	1	0	1	1	1	**5**	0	1	26	42
	2-3 times daily	0	0	1	1	1	3	2	**5**	1	7	21
	>3 times daily	1	1	1	0	0	0	0	2	**4**	13	22
	Missing	1	0	0	0	0	0	0	0	0	3	4
	Total number	514	26	27	28	33	20	15	9	9	307	988

Table 4. Cross frequency table of termination status (MATS vs. Record Abstract)

	Row %	Record Abstract Data									
		Completed treatment	Referred	W/o clinic agreement	Died	Incar-cerated	Moved/ cannot attend	Non-compliance /refused treatment	Dismissed due to program cut/reduction	Missing	Total
MATS	Completed treatment	**299**	24	12	0	0	5	7	0	6	353
	Referred	35	**29**	10	0	0	2	8	0	1	85
	W/o clinic agreement	16	6	**283**	0	3	5	43	0	3	359
	Died	0	0	0	**2**	0	0	0	0	0	2
	Incarcerated	0	0	0	0	**7**	0	0	0	0	7
	Moved/ cannot attend	10	2	6	0	0	**27**	2	0	2	49
	Noncompliance / refused treatment	5	4	41	0	0	2	**59**	0	0	111
	Dismissed due to program cut/ reduction	0	2	0	0	0	0	1	**0**	2	5
	Missing	6	3	2	0	0	0	2	0	4	17
	Total number	371	70	354	2	10	41	122	0	18	988

section, we test hypotheses only after deleting those data points with missing reports. Moreover, in our sensitivity analysis, we use multiple imputation to test the robustness of our results with respect to the missing data.

Table 3 reports the joint distribution of the discharge alcohol use frequency measure in the two data sets. According to the record abstract data, a total of 514 clients achieved abstinence at time of discharge, one of the primary alcohol abuse treatment goals. For these 514 clients, 465 (90.47 percent) were also reported as achieving abstinence in MATS. Nevertheless, for the rest of the table, a high degree of inconsistency between the MATS and record abstract data is observed. For example, out of the 28 clients reported as drinking once per week in the record abstract data, only 7 were reported in the same category in MATS. Again, both exaggeration and understatement of the discharge alcohol use frequency are common in MATS. Information on discharge frequency in the record abstract data is missing in more than 300 or 31.07 percent of clients.

Table 4 reports the frequency distribution of termination status. A total of 371 clients completed treatment according to the record abstract data. Of these 299 (80.59 percent) were reported similarly in MATS, while the rest were reported as being referred (9.43 percent), terminated treatment without any explanation (4.31 percent), or moved (2.70 percent). It appears that the degree of inconsistency between the MATS and record abstract reports on termination status is less compared to the admission alcohol use frequency in **Table 2**.

Table 5 presents the frequency distribution of the admission employment status reported in the MATS and record abstract data. For the 576 clients reported as not employed at time of admission in the record abstract data, 563 (97.74 percent) are reported likewise in MATS. Similarly, consistent reports in MATS are observed in more than 96 percent of the

Table 5. Cross frequency table of admission employment status (MATS vs. Record Abstract)

		Record Abstract Data			
		Not Employed	Employed	Missing	Total number
MATS	Not Employed	563	16	5	584
	Employed	13	387	1	401
	Missing	0	0	3	3
	Total number	576	403	9	988

Table 6. Cross frequency table of discharge employment status (MATS vs. Record Abstract)

		Record Abstract Data			
		Not Employed	Employed	Missing	Total number
MATS	Not Employed	485	9	16	510
	Employed	8	442	25	475
	Missing	0	0	3	3
	Total number	493	451	44	988

403 clients classified as employed in the record abstract data. There are only 9 clients with missing admission employment status in the record abstract data; 3 in MATS. **Table 6** shows the frequency distribution of discharge employment status in the two data sets. As in Table 5, the degree of consistency is high, although the number of missing reports is a little higher. In summary, the MATS and record abstract data exhibit a higher degree of consistency on employment status than on alcohol use frequencies and termination status.

Data Analytic Procedures

The MATS and record abstract reports on alcohol use frequencies and termination status do not appear to be consistent. While the casual observations are informative, we would like to test for these differences in a formal and statistical way. Ruling out that the inconsistencies are due to random errors, we then can conclude that measures reported in the MATS and record abstract data are systematically inconsistent. We perform chi-square tests based on contingency tables to test our null hypothesis that the observed differences are due to random errors. We also use Kappa statistics as well as sensitivity and specificity tests to support our results.

We begin by examining admission alcohol use frequency. This is a discrete variable with nine categories. Let the index i =1,2 denote whether a data point was reported in MATS or record abstract data, and index j = 1,2,…,9 denote the nine categories. The following two-by-nine contingency table is constructed by putting MATS and record abstract data along the rows and admission alcohol use frequency along the columns. The number in each cell of the contingency table, n_{ij}, refers to the total count of clients reported under alcohol use frequency j in data i:

	j = 1 not drinking in the past 30 days	j=2 once per month	…	j =9 More than three times daily
i =1 MATS	n_{11}	n_{12}	…	n_{19}
i =2 Record Abstract Data	n_{21}	n_{22}	…	n_{29}

Our Chi-square test is based on the above contingency table, not the cross-frequency table (**Table 2**). The null hypothesis is that the distributions of admission alcohol use frequency in MATS and record abstract data are identical, i.e. the probability of admission alcohol use frequency being reported as j is independent of whether the report is from MATS or record abstract data. ***** We implement the Chi-square test to examine whether the deviations of the observed counts in each cell from the expected counts are too large to be attributable to chance under the null hypothesis. **†** The larger is the Chi-square statistic, the larger the differences in the distributions, and therefore the stronger the evidence against the null hypothesis. The Chi-square test P-value indicates at what level of significance will the null hypothesis be rejected.

Besides the Chi-square test, we calculate the Kappa statistic to give a quantitative measure of the association between the admission alcohol use frequency reported in the two data sets. Kappa statistic compares the observed consistency to expected consistency by chance if the two measures were independent.[23] When measures in the two data sets are perfectly consistent, the Kappa statistic is 1. The less consistent they are, the lower the value of the Kappa, which becomes 0 when the two measures are independent. Following Rosner,[23] we interpret any Kappa statistic lower than 0.40 as indicating a poor level of consistency; higher than 0.75, excellent consistency.

A third method we use to test the consistency between MATS and record abstract data is sensitivity and specificity test. We redefine admission alcohol use frequency as being a heavy drinker (with admission alcohol use frequency higher than two to three days per week) or not. With record abstract data as a benchmark, sensitivity indicates the probability of a client being reported as heavy drinker in MATS given that the client is reported as heavy drinker in the record abstract data; specificity indicates the probability of a client being reported as not a

* The null hypothesis of a Chi-square test based on **Table 2** is that there is no association between the admission alcohol use frequency reported in MATS with that reported in the record abstract data. It is clear that MATS and the record abstract data measures are associated. This is not something we want to test. Rather, we want to test whether such association is low.

† Let row totals be denoted by N_i and column totals by M_j, grand total by N. The Pearson Chi-square statistic is given by:

$$\chi^2 = \Sigma \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}, \mu_{ij} = \frac{N_i * M_j}{N}$$

Table 7. Consistency test results on five measures

	Sample Size	Chi-square Test [1,6] (χ^2)	Kappa Statistic[2,6] (k)	Sensitivity[3,7]	Specificity[4,7]
Admission Alcohol Use Frequency [5]	696	χ^2 = 129.60*** (DF = 8) Null Hypothesis Rejected	k = 0.25*** Poor consistency	0.76 (0.71,0.81)	0.77 (0.73,0.81)
Discharge Alcohol Use Frequency [5]	680	χ^2 = 2.62 (DF = 8) Null Hypothesis not rejected	k = 0.48*** Good consistency	0.64 (0.50,0.77)	0.96 (0.94,0.97)
Termination Status [5]	957	χ^2 = 6.68 (DF=7) Null Hypothesis not rejected	k = 0.63*** Good consistency	0.82 (0.78,0.86)	0.92 (0.89,0.94)
Admission Employment Status	979	χ^2 = 0.02 (DF=1) Null Hypothesis not rejected	k = 0.94*** Excellent consistency	0.96 (0.94,0.98)	0.98 (0.96,0.99)
Discharge Employment Status	944	χ^2 = 0.00 (DF=1) Null Hypothesis not rejected	k = 0.96*** Excellent consistency	0.98 (0.96,0.99)	0.98 (0.97,0.99)

Notes:
1. Null hypothesis for the Chi-square tests: the distributions of the tested measure in MATS and record abstract data are identical; all observed inconsistencies are due to random error.
2. Guidelines for the evaluation of Kappa statistic[23]:
0.00-0.40 Poor consistency
0.41-0.75 Good consistency
0.76-1.00 Excellent consistency
3. Sensitivity = Pr (MATS = YES | Abstract Record = Yes)
4. Specificity = Pr (MATS = No | Abstract Record = No)
5. Since sensitivity and specificity tests could only be performed on binary variables, the first three measures are re-defined to obtain the sensitivity and specificity test results. Admission and discharge alcohol use frequencies are re-defined as being moderate or heavy drinker at admission and discharge; termination status is re-defined as completing treatment or not at time of discharge.
6. Under Chi-square test and Kappa statistic results, *** significant at 1%; ** significant at 5%; * significant at 10%.
7. Under sensitivity and specificity test results, confidence interval of each test is reported in the parentheses.

heavy drinker given that the client is reported as not a heavy drinker in the record abstract data.

The Chi-square test, Kappa statistic, and sensitivity and specificity test results on admission alcohol use frequency are presented in the first row of **Table 7**. Excluding clients with missing admission alcohol use frequency report in either MATS or record abstract data results in a sample of 696 clients. The Chi-square test yields a test statistic of 129.60 (with 8 degrees of freedom: DF = 8), and p < 0.0001. The null hypothesis is rejected at 1% significance level. For robustness, we repeat the Chi-square test on a different definition of "consistency." Now reports are regarded as consistent when the drinking frequency reported in MATS is no more than one category higher or lower than that reported in record abstract data. The null hypothesis remains rejected at 1% significance level. * We found a highly significant Kappa statistic, but its low value of 0.25 shows poor consistency between the MATS and record abstract report on admission alcohol use frequency. With the record abstract as a benchmark, MATS report on being a heavy drinker or not at admission is only 76 percent sensitive and 77 percent specific. All three results suggest that the MATS and record abstract data reports on admission alcohol use frequency are statistically inconsistent.

The results on discharge alcohol use frequency is reported in the second row in **Table 7**. After deleting data points with missing information, we have a sample of 680. The Chi-square test statistic is 2.62 (DF = 8), and p = 0.96. The conventional threshold of rejecting the null hypothesis in standard statistical analysis is a P-value of 0.10. Under such a threshold, the null hypothesis is not rejected. It follows that if the definition of report consistency is relaxed as it is done for admission use frequency, the null hypothesis will remain unrejected. The Kappa statistic is 0.48, indicating good consistency, although it is still quite close to the threshold of 0.40 for poor consistency.[23] Heavy drinker or not at discharge reported in MATS is only 64 percent sensitive, but 96 percent specific. Our results on discharge alcohol use frequency reports are not conclusive. Further explorations on this measure can be found in the next subsection.

Next we test whether termination status reported in MATS is significantly different from the record abstract data. As we see in the third row of **Table 7**, the effective sample size is 957. A Chi-square test yields a test statistic of 6.68, and p = 0.46 (DF = 7). As before, the threshold of rejecting the null hypothesis is 0.10. The null hypothesis cannot be rejected.

* In fact, only when we relax the definition of "consistency" to including drinking frequency reported in MATS no more than four categories higher or lower than in record abstract did we get insignificant results. This indicates that our results on systematic misreporting is highly robust.

Table 8. Multiple imputation consistency test results

	Effective Sample Size	Multiple Imputation (M=5)		Multiple Imputation (M=10)	
		Chi-square test[1,2,5] (t: combined Chi-square statistic)	Kappa Statistic[3,4,5] (k: combined Kappa statistic)	Chi-square test[1,2,5] (t: combined Chi-square statistic)	Kappa Statistic[3,4,5] (k: combined Kappa statistic)
Admission Alcohol Use Frequency	988	t = 175.99*** (DF=5.11) Null Hypothesis Rejected	k = 0.19*** (DF=38.62) Poor consistency	t = 174.12*** (DF=13.50) Null Hypothesis Rejected	k = 0.19*** (DF=102.45) Poor consistency
Discharge Alcohol Use Frequency	988	t = 22.64*** (DF=7.56) Null Hypothesis Rejected	k = 0.41*** (DF=419.90) Poor~good consistency	t = 25.23*** (DF=34.88) Null Hypothesis Rejected	k = 0.42*** (DF=318.85) Poor~good consistency

1. Null hypothesis for the Chi-square tests: the distributions of the tested measure in MATS and record abstract data are identical; all observed inconsistencies are due to random error.
2. The combined Chi-square statistic t, which is the mean of all chi-square test statistics of each imputed data set, follows a student t-distribution with an adjusted degree of freedom.
3. The combined Kappa statistic K, which is the mean of all Kappa statistic of each imputed data set, follows a student t-distribution with an adjusted degree of freedom.
4. Guidelines for the evaluation of Kappa statistic,[23]
   0.00-0.40 Poor consistency
   0.41-0.75 Good consistency
   0.76-1.00 Excellent consistency
5. *** significant at 1% significance level; ** significant at 5% significance level; * significant at 10% significance level.

The Kappa statistic is highly significant, and has a value of 0.63. Following guidelines for evaluation of Kappa statistic,[23] this indicates a degree of consistency higher than alcohol use frequency measures, but not excellent. Redefining the termination status as a binary variable indicating treatment being complete or not, we find that the MATS report is 82 percent sensitive and 92 percent specific. There is mixed evidence on report consistency on termination status in the two data sets.

Finally, we test for the consistency between MATS and record abstract data on employment status. The results on admission employment status are in the fourth row of **Table 7**. With a sample size of 979, the Chi-square statistic is 0.02 (DF = 1), with p = 0.89. The null hypothesis cannot be rejected. The Kappa statistic is 0.94, indicating significant and excellent consistency. Using admission employment status reported in record abstract data, the MATS report is found to be 96 percent sensitive and 98 percent specific. These results suggest that admission employment status reported in the two data sets are highly consistent. Similarly, client discharge employment status reports are found to be highly consistent. The sample size for this test is 944; see the last row of **Table 7**. A Chi-square test yields a test statistic of 0.00 (DF = 1), and p = 0.96.***** The Kappa statistic is 0.96 and highly significant, and MATS report is 98 percent sensitive and 98 percent specific.

According to the tests, MATS evaluations and abstracted records are systematically and significantly different with respect to admission alcohol use frequency, but not termination and employment status. Results on discharge alcohol use frequency are less conclusive. Furthermore, reports in the two data are more consistent with respect to employment status than termination status.

Sensitivity Analyses

A major concern on the alcohol use frequency test results is the approximately a third of missing values in the record abstract data. By contrast, record abstract data reports on employment status are only missing less than 4.5 percent of the sample; termination status, less than two percent. How reliable are the above test results on alcohol use frequencies when we could only use two third of the sample? In this section, we employ multiple imputation methods to address robustness issues.

Multiple imputation is a statistical approach developed in recent decades to address missing data problems.[24-27] This approach involves "imputing" M values for each missing item in the data set and creating M completed data sets. The purpose of having many imputations is to account for the uncertainty under which the missing data are imputed from the observed data. Each completed data set can be analyzed by any standard method. For our study, for each imputed data set, we test whether the admission and discharge alcohol use frequencies are consistent with the Chi-square test and Kappa statistic. Finally, all M Chi-square test statistics and Kappa

***** Fisher's Exact tests were also performed to test gaming on admission and discharge employment status. The results are consistent with those from the Chi-square tests.

statistics are combined using rules from Rubin.[25]

A necessary condition of using multiple imputation is that the missing data must be missing at random (MAR) as defined by Rubin.[28] Let D denote the data matrix, and D^{obs} and D^{mis} respectively the observed and missing components. MAR assumes that the distribution of D^{mis} is only dependent on D^{obs}.[28, 29]

In our study, the data matrix D includes the admission and discharge alcohol use frequencies from both the record abstract and MATS. In addition, the following MATS variables are also included in D: admission and discharge employment status, age, sex, marital status, education, legal involvement at admission, concurrent psychiatric problem, household income, severity of substance abuse at admission, number of prior treatment episodes, primary payer status, whether the client was admitted after PBC has been implemented, and whether the client was discharged after PBC was introduced.

The majority of the missing elements D^{mis} are the missing record abstract admission and discharge alcohol use frequencies. As we have discussed earlier, the sources of record abstract data are hand-written clinical records. There is no fixed format on how alcohol use frequency should be documented in the clinical records. The missing data result largely from a cautious data collection methodology. In addition, we have included in the data matrix a range of variables on client characteristics, alcohol use severity, insurance sources, and contracting system. It is plausible to assume that the missing data in our sample are due to chance after conditioning on D^{obs}.

The above MAR assumption ensures the feasibility of the imputation. The goal of the imputation stage is to generate random draws on missing values in D from some distribution $f(D^{mis}|D^{obs})$. In this study, we adopt the Markov Chain Monte Carlo (MCMC) approach, which is often recommended for missing-at-random data problems.[29,30] Rather than drawing directly from $f(D^{mis}|D^{obs})$, we use a Markov chain, $\{D^{mis(1)}, D^{mis(2)}, \ldots, D^{mis(t)}, \ldots\}$, to simulate draws. The distribution of each element in the Markov chain depends on the value of the previous one. The distribution of $D^{mis(t)}$ as t goes to infinity is $f(D^{mis}|D^{obs})$. When t is sufficiently large, $D^{mis(t)}$ is approximately a random draw from f. Various computation algorithms are available to implement MCMC. In this paper, we present results from the Imputation-Posterior (IP) algorithm.[29] *

Although MCMC approaches require a multivariate normality assumption, inferences from this approach are shown to be robust to minor departures from such an assumption.[29] Since both admission and alcohol use frequencies are categorical variables (with nine categories), we use log transformation in the imputations to avoid large violation of the multivariate normality assumption.

In the next step, using each of the M "completed" data sets, we test whether reports on admission and discharge alcohol use frequencies are consistent. Let Q_i and U_i denote the Chi-square test (or Kappa) statistic and variance estimate from the i^{th} data set. According to Rubin's rule for combining imputation results, the multiple imputation estimate of the Chi-square test statistics is simply the average of the M separate Chi-square test statistics (Q_is); likewise for the Kappa statistic. Each of these multiple imputation estimate follows a student t distribution with an adjusted degree of freedom.[32] *

Only a small number of imputations is necessary to obtain accurate and valid inferences. According to Rubin,[25] with 30 percent missing information, an estimate based on M = 5 or 10 imputations will tend to have a standard error only 1.03 or 1.01 times as wide as that based on an infinite number of imputations. So we performed 5 and 10 imputations. Multiple imputation Chi-square test and Kappa statistics results on admission and discharge alcohol use frequencies are presented in **Table 8**. † For both cases, the Chi-square tests reject the null hypothesis that MATS and abstract record data sets are consistent. As before, we repeat the Chi-square tests after relaxing the definition of report consistency in admission and discharge alcohol use frequencies. The null hypotheses continue to be rejected. Kappa statistic shows poor consistency on admission alcohol use frequency (0.19), and poor to good consistency on discharge alcohol use frequency (0.41). The imputation results support our claims in **Table 7** on admission alcohol use frequency, and suggest significant inconsistency in discharge alcohol use frequency reports.

Concluding Remarks

In this paper, we investigate information consistency across two data sets. Medical record abstract data are compared to an administrative data set, Maine Addiction Treatment System. The comparison is on a clinician's reports, as recorded by these two data sets, on an alcohol abuse treatment episode. Admission and discharge alcohol use frequencies reported in MATS are shown to be significantly different from the record abstract data. Nevertheless, reports on admission and discharge employment status are found to be highly consistent in these two data sets. There is mixed evidence on report consistency on termination status. Sensitivity analyses affirm the robustness of the above results. It will be of interest to confirm our finding with other data sets and settings.

* The software we use is AMELIA, a Gauss-based program developed by Gary King at Harvard University[31] We also used Expectation-Maximization (EM), EM with sampling (Ems), and EM with importance resampling (Emis) algorithms provided by AMELIA. The results of consistency tests when using these algorithms are consistent with what presented in the paper.

* According to Schafer and Olsen,[32] the degree of freedom of this student t distribution, df, is defined by:

$$df = (M-1)[1 + \frac{M \cdot \bar{U}}{(M+1)B}]^2$$

$$\bar{U} = \frac{1}{M}\sum_{i=1}^{M} U_i$$

$$B = \frac{1}{M-1}\sum_{i=1}^{M}(Q_i - \bar{Q})^2$$

† Very few admission and discharge employment status are missing in the record abstract data. Therefore, we do not perform multiple imputations on admission and discharge employment status variables.

Our work in this paper on the data sets confirms the existence and statistical significance of strategic reporting in alcohol addiction treatment. The effect of the report inconsistency on managed-care practice as well as care delivery cannot be studied using only the data sets we have used. The information of the actual insurance and health-care policy for the covered population will be necessary for this analysis; a larger scale study is needed.

Our on-going research will model the motives behind strategic reporting. We will hypothesize that both altruistic and financial incentives are present. Our empirical identification strategy will use Maine's Performance-Based Contracting system and client insurance sources to test how these incentives affect the direction of clinician's strategic reporting. Our preliminary work supports the hypothesis that financial incentives significantly affect the direction of report inconsistency.

Acknowledgement

The authors thank the National Institute on Alcohol Abuse and Alcoholism. Lu also gratefully acknowledges financial support from the Alberta Heritage Foundation for Medical Research and the Institute of Health Economics. At various stages of this project, Priyodorshi Banerjee, Lee McGuire, Jitian Sheu, and Jenny X. Zhang provided excellent research assistance. For their comments and suggestions, the authors thank Ake Blomqvist, John Bridges, Qing J. Liang, Hisen-Ming Lien, Matthew Maciejewski, Thomas G. McGuire, Sharon-Lise Normand, LieLing Wu, and seminars participants at the 77th Annual International Conference of the Western Economic Association as well as at the University of Hong Kong. The authors alone are responsible for the analysis and conclusions.

References

1. Morreim HE. Gaming the system: dodging the rules, ruling the dodgers. *Arch Intern Med* 1991; **151**: 443-447.
2. Cain JM. Is deception for reimbursement in obstetrics and gynecology justified? *Obstet Gynecol* 1993; **82**: 475-478.
3. Meskin LH. The noble lie. *The Journal of American Dental Association* 2000; **131**: 556-560.
4. Novack DH, Detering BJ, Arnold R, et al. Physicians' attitudes toward using deception to resolve difficult ethical problems. *JAMA* 1989; **261**: 2980-2985.
5. Carter MG, Newhouse JP, Relles DA. *How Much Change In the Case Mix Index is DRG Creep?* Working paper. Santa Monica: RAND.
6. Commons M, McGuire TG. 1997. Some Economics of Performance-Based Contracting for Substance-Abuse Services. In *Treating Alcohol Abusers Effectively*, Egertson, JA, Fox DM, Leshner AI, (eds). Milbank Memorial Fund and Blackwell Publishers, 223-249.
7. Lu M. Separating the "True Effect" from "Gaming" in Incentive-Based Contracts in Health Care. *Journal of Economics and Management Strategy* 1999; **8**: 383-432.
8. Carter MG, Newhouse JP, Relles DA. *Has DRG Creep Crept Up? Decomposing the Case Mix Index Change Between 1987 and 1988.* Working paper. Santa Monica: RAND 1991.
9. Jonkman JN, Normand SLT, Wolf R, Borbas C, Guadagnoli E. Identifying a Cohort of Patients with Early-Stage Breast Cancer: A Comparison of Hospital Discharge and Primary Data. *Med Care* 2001; **39**: 1105-1117.
10. Warren JL, Riley GF, McBean AM, Hakim R. Use of Medicare data to identify incident breast cancer cases. *Health Care Financ Rev* 1996; **18**: 237-246.
11. McLish DK, Penberthy L, Whittemore M, Newschaffer C, Woolard D, Desch CE, Retchin S. Ability of Medicare claims data and cancer registries to identify cancer cases and treatment. *Am J Epidemiol* 1997; **145**: 227-233.
12. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. The sensitivity of Medicare claims data for case ascertainment of six common cancers. *Med Care* 1999; **37**: 436-444.
13. Warren JL, Feuer E, Potosky AL, Riley GF, Lynch CF. Use of Medicare hospital and physician data to assess breast cancer incidence. *Med Care* 1999; **37**: 445-456.
14. Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. *J Clin Epidemiol* 2000; **53**: 605-614.
15. Solin LJ, Legorreta A, Schultz DJ, Levin HA, Zatz S, Goodman RL. Analysis of a claims database for the identification of patients with carcinoma of the breast. *J Med Syst* 1994; **18**: 23-32.
16. Leung KM, Hasan AG, Rees KS, Parker RG, Legorreta AP. Patients with newly diagnosed carcinoma of the breast: validation of a claim-based identification algorithm. *J Clin Epidemiol* 1999; **52**: 57-64.
17. Cooper GS, Yuan Z, Stange KC, Dennis LK, Amini SB, Rimm AA. Agreement of Medicare claims and tumor registry data for assessment of cancer-related treatment. *Med Care* 2000; **38**: 411-421.
18. Du X, Freeman JL, Warren JL, Nattinger AB, Zhang D, Goodwin JS. Accuracy and completeness of Medicare claims data for surgical treatment of breast cancer. *Med Care* 2000; **38**: 719-727.
19. Maine Office of Substance Abuse. *Maine Addiction Treatment System Instruction Manual*. Augusta, Maine, June 1994.
20. Lu M, McGuire TG. The Productivity of Outpatient Treatment for Substance Abuse. *J Hum Resour* 2002; **37**:309-355.
21. Commons M, McGuire TG, Riordan MH. Performance Contracting for Substance Abuse Treatment. *Health Serv Res* 1997; **32**: 631-650.
22. Lu M, Ma CtA, Yuan L. Risk Selection and Matching in Performance-Based Contracting. *Health Econ*, forthcoming.
23. Rosner B. *Fundamentals of Biostatistics*. Fifth edition. Pacific Grove: Duxbury, 2000.
24. Rubin DB. Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *J Am Stat Assoc* 1977; **72**: 538-543.
25. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: Wiley, 1987.
26. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; **91**: 473-489.
27. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.
28. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**: 581-592.
29. Schafer JL. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.
30. Horton NJ, Lipsitz SR. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables, *Am Stat* 2001; **55**: 244-254.
31. King G, Honaker J, Joseph A, Scheve K. Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* 2001; **95**: 49-69.
32. Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research* 1998; **33**: 545-571.