How valid is self-reported health data?

A Chapman Institute White Paper

May 2012
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface</td>
<td>3</td>
</tr>
<tr>
<td>Executive Summary</td>
<td>4</td>
</tr>
<tr>
<td>A. Problem or Issue</td>
<td>6</td>
</tr>
<tr>
<td>B. Summary of Literature Review Findings</td>
<td>9</td>
</tr>
<tr>
<td>C. Chapman Institute Position</td>
<td>15</td>
</tr>
<tr>
<td>D. Implications for Programming</td>
<td>17</td>
</tr>
<tr>
<td>E. Selected Abstracts with Highlighted Summaries</td>
<td>18</td>
</tr>
<tr>
<td>F. Glossary of Terms</td>
<td>41</td>
</tr>
<tr>
<td>G. Bibliography on the Validity and Reliability of HRAs</td>
<td>49</td>
</tr>
</tbody>
</table>
Preface

As one of the premier health management companies in the United States, Chapman Institute has provided consulting services to hundreds of clients with a wide range of health and wellness program design characteristics. In an effort to build the body of knowledge underpinning health management Chapman Institute provides technical monographs analyzing topical issues within the discipline. These *White Papers* are designed to assist clients, fellow practitioners and other interested parties in pursuing technical excellence in health management programming.

**Objectives of the Chapman Institute monograph series:**

- To present technical insights into the potential solution to a central question or challenge facing those who conduct health management programming.
- To establish the basis of justification for a technical position or policy that Chapman Institute uses to guide the design and operation of its products and services.
- To further discussion and consensus on important issues affecting the field of wellness and health management.
- To present technical issues in ways that are easy for all stakeholders in health management to understand.

This monograph has been prepared by Chapman Institute’s President and CEO Larry S. Chapman, MPH.
Executive Summary

Issue:
How valid is self-reported health data?

Chapman Institute Position:
“Based on our review of a considerable body of scientific literature, self-reported health data is valid for effective use in population health management interventions.”

Summary of Rationale:
The Chapman Institute Position is based on the following:
A large body of research has clearly documented that the average individual, and particularly the average employee (and his/her spouse) is the most accurate and valid source of information about his/her own health needs, interests, current behaviors and readiness to change. This body of scientific findings is summarized later in this document in Section E below.
This body of research confirms the validity of the use of self-reported health data in the following areas:

- Pain reporting
- Neuromuscular function
- Physical activity levels
- Height and weight
- Breast and cervical screening exams
- Alcohol consumption patterns
- Vitamin and supplement use
- Nutrition practices
- Tobacco use
- Hypercholesterolemia
• Hypertension
• Diabetes
• Influenza vaccination
• Clinical breast exam
• Mammography
• Pap smear
• Blood pressure screening
• Stroke risk factors and prevalence
• Sleep conditions and patterns
• Angina
• Ischemic heart disease
A. Problem or Issue

This monograph addresses the issue of the basic validity of self-reported health data, particularly in the context of answers provided by individuals in health risk assessments (HRAs), in paper or web versions. The summary statement of the issue under examination in this monograph is identified in Figure 1:

**ISSUE**: Is self-reported health data valid? In other words, how truthful and/or knowledgeable are individuals when they complete HRAs? And even more importantly... can the data they provide be effectively used for purposes of health management planning, individual intervention and evaluation?

Figure 1: Summary Statement of the Issue

Validity and Reliability of Health Risk Assessment Data:
Annual health risk assessments (HRAs) are a core element of all proactively-oriented worksite wellness and health promotion programs.¹ Over the past 35 years, a great deal of development has been done on the validity and reliability of HRAs. Dee W. Edington, Louis Yen, and Alexandra Braunstein have conducted some of the most insightful analysis on these topics for publication in the *Handbook of Health Assessment Tools*, which was published by the Society of Prospective Medicine and the Institute for Health and Productivity Management.²

---

The highlights of their findings include:

- Early HRA validity and reliability was improperly assessed in widely varying applications.
- Unrealistic expectations for predictive accuracy clouded assessments of HRA utility.
- HRAs are very effective as awareness, education and risk identification tools.
- HRAs have more limited value as predictors of risk, morbidity or mortality at the individual level.
- When appropriately designed, HRAs can be tools for prediction of group level risks, utilization and cost.
- The general scientific agreement is that HRAs have a high degree of face validity.
- HRA responses are consistently reliable with reliability coefficients in the range of .75 to .99, depending on the types of question asked.
- An overly negative assessment of reliability in early studies on HRA reliability did not adequately adjust for programming effects between administrations of HRAs.
- HRA questions accurately represent what they were designed to do.
- HRA completers are generally healthier than non-completers unless the completion rate for the HRA in the population is extremely high (greater than 70%).
- HRAs are accurate predictors of three year health costs.
• HRAs are accurate in putting people into risk categories but not scientifically accurate in predicting individual risk of dying.
• HRA use is a cost-effective way to identify medical problems.
• HRAs are more cost-effective than are many of the more traditional elements of the patient care database and often detect problems that other elements of the database miss.
• HRAs increase self-efficacy.

The prevention and health promotion fields have reached a broad-based consensus in the use and validity of HRA data in health management. In addition, the emerging paradigm of Health and Productivity Management (HPM) is providing impetus to further expand the role and function of the HRA. Expansion of HRA question sets and custom text block responses to address sick leave, workers’ compensation, disability management and presenteeism appears close. This expanded function provides an even more robust strategic rationale to build HRAs into annual benefit processes and to link interventions to the information provided in HRAs to planning, targeting and evaluation activities.

---

B. Summary of Literature Review Findings

Chapman Institute retrieved and selected 19 peer review articles through the literature search process using a formal set of selection or inclusion parameters. Chapman Institute’s process for the literature search and the accompanying analyses are described below.

Literature Search Parameters

Parameters for the article search process included:

1. Must be in a peer review publication.
2. Must be in English.
3. Must be in the United States, Canada, Mexico or United Kingdom due to the cultural congruence with United States employee and spousal populations and issues of external validity.
4. Must address the issue of the validity of self-reported health data.
5. Must use statistical analysis techniques on the observed change.
6. Must be original research results.
7. Must be published after 1990.
8. Must be included in the study pool if they meet the above criteria.

Outline of Literature Search Methods

The Medline search engine of the National Library of Medicine was used to find applicable articles and studies. The search method included the use of strings of key search terms across the full Medline research data base. The bibliographies of each of the retrieved articles were back searched for additional relevant articles. Chapman Institute uncovered 19 peer review articles in the literature search process.
Summary of Literature Review Findings

Each of the peer review articles is highlighted in Table 1 below. The table contains the citation number of the study which are presented based on their publication date with the most recent first, a description of the characteristics of the study population, the actual language of the principal conclusion from the study regarding the validity of self-reported health information and additional comments from the study. See above for the study inclusion criteria.
Table 1:  Summary of Literature Review Findings

<table>
<thead>
<tr>
<th>Citation #</th>
<th>Study Population Characteristics (Study Population Size)</th>
<th>Study Conclusions</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>Adults with neuromuscular disorders (NMD) and chronic pain (N = 141)</td>
<td>“The FIM-SR (Self-Report) scales appear to be reliable and valid measures of independence in 6 specific (self-care, sphincter control, mobility, locomotion, communication, social cognition) and 3 global (motor, cognition, total) areas of functioning in persons with NMD.”</td>
<td>Conducted in rehabilitation population</td>
</tr>
<tr>
<td>#2</td>
<td>National sample of Medicare beneficiaries (N = 6,538)</td>
<td>“In the derivation cohort, 14 self-reported characteristics were significant predictors of developing a need for long-term care within 1 year.”</td>
<td>Self-reported characteristics included activity levels, pain levels, functional capabilities, activities of daily living and psycho-social variables.</td>
</tr>
<tr>
<td>#3</td>
<td>National sample of Mexican adult citizens (N = 1,707)</td>
<td>“Height and weight self-report is a valid method that may be used to accurately estimate height and weight in Mexican people.”</td>
<td>Error increases with age and females tend to under-report weight and over-report height more than males.</td>
</tr>
<tr>
<td>#4</td>
<td>National sample of obese adults age 18 to 75 (N = 6,107)</td>
<td>“The Obesity and Weight Loss Quality of Life and Weight Related Symptoms Measures are brief, valid, reproducible, and responsive self-reported outcomes for evaluating obesity and weight loss.”</td>
<td>Tests of self-reported psychometric and biometric surveys with obese subjects.</td>
</tr>
<tr>
<td>#5</td>
<td>Adult women age 40 to 75 in a managed care setting (N = 480)</td>
<td>“These results suggest that self-reporting of breast and cervical cancer screening is fairly accurate in this managed care population, although women tend to underestimate the time since their last screening.”</td>
<td>Error increases with age.</td>
</tr>
<tr>
<td>#6</td>
<td>Large sample of male and female Army active duty personnel (N = 404,966)</td>
<td>“The Army's HRA alcohol items seem to elicit reliable and valid responses. Because HRAs contain identifiers, alcohol use can be linked with subsequent health and occupational outcomes, making the HRA a useful epidemiological research tool.”</td>
<td>Comparisons with subsequent alcohol-related morbidity and administrative actions were used to validate self-report data.</td>
</tr>
<tr>
<td>#7</td>
<td>Sample of adult males and females (N = 220)</td>
<td>“This self-administered questionnaire demonstrated high reproducibility and validity for collecting detailed</td>
<td>Comparisons with blood and urine samples and label</td>
</tr>
<tr>
<td>Citation #</td>
<td>Study Population Characteristics (Study Population Size)</td>
<td>Study Conclusions</td>
<td>Comments</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------------------</td>
<td>--------------------</td>
<td>----------</td>
</tr>
<tr>
<td>#8</td>
<td>National sample of male and female adults age 35 to 76 (N = 4,808)</td>
<td>“Self-reported height and weight data are valid for identifying relationships in epidemiological studies. In analyses where anthropometric factors are the primary variables of interest, measurements in a representative sample of the study population can be used to improve the accuracy of estimates of height, weight and BMI.”</td>
<td>Height error greater in men than women and increased with age. Weights are underestimated an average of 1.85 kg in men and 1.40 kg in women leading to a misclassification in BMI of 22.4% of men and 18.0% in women</td>
</tr>
<tr>
<td>#9</td>
<td>National sample of male and female adults (N = 8,326)</td>
<td>“Due to low sensitivity, self-reported hypercholesterolemia should be used with caution, both during the patient encounter and for surveillance of trends in hypercholesterolemia in the absence of measured cholesterol levels. Specificity is consistently much higher than sensitivity. The high Positive Predictive Value (PPV) may be of use in certain clinical situations. Such validation studies should form the foundation for future research based on self-report.”</td>
<td>Sensitivity was higher with older individuals and non-Hispanic whites. Specificity was greater with individuals with more than 12 years of education. PPV was greater in older persons.</td>
</tr>
<tr>
<td>#10</td>
<td>Nationally representative sample of the US civilian, non-institutionalized population (N = 16,573)</td>
<td>“Self-reported heights and weights can be used with younger adults, but they have limitations for older adults, ages &gt; or = 60 years.”</td>
<td>Age is an important factor in classifying weight, height, BMI, and overweight from self-reports.</td>
</tr>
<tr>
<td>#11</td>
<td>An epidemiological cohort of adult males and females age 56 to 78. (N = 3,182)</td>
<td>“Self-reported measures of height and weight may be used in studies of the elderly although systematic reporting errors may bias effect estimates.”</td>
<td>The overweight under-reported their weight to a greater extent.</td>
</tr>
<tr>
<td>Citation #</td>
<td>Study Population Characteristics (Study Population Size)</td>
<td>Study Conclusions</td>
<td>Comments</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>#12</td>
<td>A random, stratified sample of male and female adults &gt; 21 years of age in an HMO population in Colorado (N = 550)</td>
<td>“Self-reports are reasonably accurate for certain chronic conditions and for routine screening exams and can provide a useful estimate for broad measures of population prevalence.”</td>
<td>Three chronic conditions and six preventive services were studied by comparing self-report to clinical records.</td>
</tr>
<tr>
<td>#13</td>
<td>A national random, stratified sample of males and females from the National Health Survey for an epidemiological follow-up study. (N = Unknown)</td>
<td>“These results furnish some evidence that analyses based on self report can provide valid, useful information.”</td>
<td>Focus of this epidemiological follow-up study was on stroke risk factors and stroke prevalence.</td>
</tr>
<tr>
<td>#14</td>
<td>Convenience sample of 18-45 year old males and females in the Detroit area (N = 2,181)</td>
<td>“This study replicated the results of the clinic-based study and suggested a potentially useful diagnostic threshold for self-report excessive daytime sleepiness. Epidemiology of sleep depends on the ability to move from the laboratory to population surveys in reliable and valid ways. Development of self-report is a step in that direction.”</td>
<td>Split scales were used to test reliability and validity of self-reported sleepiness to enable the condition to be examined outside of the clinical laboratory with working populations.</td>
</tr>
<tr>
<td>#15</td>
<td>A national sample of men in the United Kingdom (N = 5,789)</td>
<td>“These results suggest that self-reported history of a doctor diagnosis of angina is a valid measure of diagnosed angina in population-based studies in men.”</td>
<td>National sample of men compared to self-reported doctor defined angina with medical records over a 12 year period.</td>
</tr>
<tr>
<td>#16</td>
<td>A low income urban population using a teaching hospital as their site of care. (N = 477)</td>
<td>“The accuracy of self-report of mammography and Pap smear is relatively poor for medical practice but is acceptable in population surveys with appropriate correction for over-reporting.”</td>
<td>Low income women tend to over-report the occurrence of Pap tests and mammography.</td>
</tr>
<tr>
<td>Citation #</td>
<td>Study Population Characteristics (Study Population Size)</td>
<td>Study Conclusions</td>
<td>Comments</td>
</tr>
<tr>
<td>------------</td>
<td>--------------------------------------------------------</td>
<td>-------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>#17</td>
<td>A national, randomized, stratified, sample of the National Health Survey (N = Unknown)</td>
<td>“Self-reported hypertension may be used for surveillance of hypertension trends, in the absence of measured blood pressure, among non-Hispanic whites and non-Hispanic black women and persons with a medical visit in the past year.”</td>
<td>Validity of self-reported hypertension is higher among women than among men and among persons with a medical visit during the past year than among those with no visits: validity was lowest among Mexican-American men.</td>
</tr>
<tr>
<td>#18</td>
<td>A provincial sample from Manitoba, Canada comparing self-report with medical records and administrative data. (N = 2,792)</td>
<td>“This study demonstrates the feasibility of linking a large health survey with administrative data and the validity of self-reports in estimating the prevalence of chronic diseases, especially diabetes and hypertension.”</td>
<td>For diabetes and hypertension, self-reports were fairly accurate in detecting &quot;true&quot; past history of the illness based on physician diagnosis recorded on insurance claims.</td>
</tr>
<tr>
<td>#19</td>
<td>An isolated military base was used as the study population. (N = Unknown)</td>
<td>“It is concluded that questions that produce valid and reliable responses do so for identifiable reasons, and measurement instruments can be improved by incorporating particular features.” (Alcohol behaviors)</td>
<td>The results of eight measures of consumption of alcohol were compared with apparent consumption, as established by documented sales, and the validity and reliability of the various measures were determined using the classical correlational approach.</td>
</tr>
</tbody>
</table>

Note: These studies reflect articles with the terms “self-report” and “validity” in their title. These search parameters are believed to provide a more accurate and efficient search process in locating articles whose purposes are aligned more closely with the purpose of this White Paper.
C. Chapman Institute Position

Chapman Institute Position:

“Based on our review of a considerable body of scientific literature, self-reported health data is valid for effective use in population health management interventions.”

Summary of Rationale:

The Chapman Institute Position on this issue is based on the following: A large body of research has documented that the average individual, and particularly the average employee (and his/her spouse) is the most accurate and valid source of information about his/her own health needs, interests, current behaviors and readiness to change. Much of this literature also documents areas where particular types of health information are less valid than others. Some of the specific instances and types of data that are prone to increased levels of self-report errors are identified below in Table 2. The implications of these findings are drawn out in the following section.

Table 2: Types of Data Prone to Self-report Errors

<table>
<thead>
<tr>
<th>Citation Number (Author)</th>
<th>Study Population Characteristics</th>
<th>Type of Data and Circumstances Where Self-Report Error Occurs (Direction of Self-Report Error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#3 (Avila-Funes)</td>
<td>Males and females Over age 75</td>
<td>Height and weight Both males and females (Under-report weight and over-report height)</td>
</tr>
<tr>
<td>#5 (Caplan)</td>
<td>Women Age 45 to 70</td>
<td>Time interval since last mammography, Clinical breast exam and Pap smear All females (Under-report time interval)</td>
</tr>
<tr>
<td>#8 (Spencer)</td>
<td>All ages Men Women</td>
<td>Height and weight Men over-estimate height by 1.23 cm on average Men under-estimate weight by 1.85 kg on average Women over-estimate height by .7 cm on average Women under-estimate weight by 1.40 kg on average. (Under and over reporting) No age effects on validity of self-report found in the study population.</td>
</tr>
<tr>
<td>Citation Number (Author)</td>
<td>Study Population Characteristics</td>
<td>Type of Data and Circumstances Where Self-Report Error Occurs (Direction of Self-Report Error)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| #9 (Natarajan)           | Males and females               | Hypercholesterolemia  
Sensitivity—knowing if they have it—is greater in older individuals and non-Hispanic whites. (Under-report)  
Specificity—knowing the actual level—is greater with those with more than 12 years of education. |
| #10 (Kuczmarski)         | Males and females               | Height, weight and BMI  
Those over 70 had a self-report error on weight and height equivalent to one BMI unit. All other ages had high correlation with self-report versus measured height and weight. (Under-report and over-report) |
| #11 (Gunnell)            | Males and females               | Height, weight and leg length  
Height over-estimated and weight under-estimated in both genders.  
Men over-estimate height more than women (2.1 cm to 1.7 cm).  
(Under-report and over-report) |
| #12 (Martin)             | Males and females in a Colorado managed care organization | Hypertension, cholesterol, diabetes, Pap smear, mammography, flu shot, clinical breast exam, cholesterol test, blood pressure screening  
Hypertension – sensitivity .83  
Diabetes – sensitivity .73  
Cholesterol – sensitivity .59  
Pap smear – sensitivity .88  
Mammography sensitivity – .91  
Flu shot – sensitivity .94  
Clinical Breast Exam – sensitivity .89  
Cholesterol test – sensitivity .86  
Blood pressure screening – sensitivity .99  
All measures had a specificity > .75  
(Under-report) |
| #17 (Vargas)             | Males and females               | Hypertension  
Men have higher self-report error than women  
Self-report error is lower if blood pressure measurement has occurred in the previous 12 months  
Highest self-report error is in Mexican men  
(Under-report) |

Note: These areas of self-report error are usually under 15% of the numerical magnitude of the health data parameter under study.

This body of scientific findings is summarized in Section E.
D. Implications for Programming

The implications of this study data are several. The most important implication is that self-reported health data of a wide range of types have been found to be valid measures for inclusion in population survey efforts. The findings from individual studies affirm the general findings from the formal research performed around the validity and reliability of health risk assessments (HRAs).

An additional implication is that small formula adjustments in specific types of health data associated with gender and age can increase validity into population survey activities. The implications translate into the following general decision rules for adjustment of HRA results in Table 3 below:

<table>
<thead>
<tr>
<th>Qualifying Participant Characteristic</th>
<th>Suggested Decision Rule Modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males over age 18</td>
<td>Subtract 1.23 cm from height measure.</td>
</tr>
<tr>
<td></td>
<td>Add 1.86 kg to weight estimate.</td>
</tr>
<tr>
<td>Males over age 70</td>
<td>Add one BMI unit to derived score.</td>
</tr>
<tr>
<td>Females over age 18</td>
<td>Subtract .7 cm from height measure.</td>
</tr>
<tr>
<td></td>
<td>Add 1.40 kg to weight estimate.</td>
</tr>
<tr>
<td>Females over age 70</td>
<td>Add one BMI unit to derived score.</td>
</tr>
<tr>
<td>Males and females</td>
<td>Utilize several types of questions to increase sensitivity and specificity of hypercholesterolemia and hypertension.</td>
</tr>
</tbody>
</table>

Note: These decision rule modifications are suggested to maximize the validity of self-report data in Health Risk Assessments (HRAs). Clearly the scientific evidence is that self-report always provides the highest predictive value.
E. Selected Abstracts with Highlighted Summaries

The following are selected abstracts on the validity of self-reported health data with highlighting in yellow of the key summary conclusion of each article. The highlighted passage has also been transferred verbatim to Table 1 above.

Citation #1: The reliability and validity of a self-report version of the FIM instrument in persons with neuromuscular disease and chronic pain.
M.P. Jensen, R.T. Abresch, and G.T. Carter

OBJECTIVE: To evaluate the reliability and validity of a self-report version of the FIM instrument (FIM-SR). DESIGN: Survey study. SETTING: Rehabilitation research program.

PARTICIPANTS: Adults with neuromuscular disorders (NMD) and chronic pain (N=141). INTERVENTIONS: Not applicable.

MAIN OUTCOME MEASURES: The FIM-SR and Medical Outcomes 36-Item Short-Form Health Survey (SF-36).

RESULTS: The internal consistency coefficients of the FIM-SR scales were all adequate to excellent (Cronbach alpha range, .73-.98; median, .96). Correlations between the FIM-SR scales and SF-36 scales supported the concurrent validity of the former. Also, the FIM-SR scales associated with motor function discriminated between those subjects who reported being ambulatory and those who reported requiring use of a wheelchair or other assistive device for getting around. Finally, FIM-SR scales discriminated between different types of NMDs, with patients with amyotrophic lateral sclerosis showing significantly lower scores on the FIM-SR self-care, motor, and total scores than all other NMD diagnostic groups, and showing significantly lower scores on the FIM-SR sphincter control, mobility, and locomotion scales than most of the
other diagnostic groups. **CONCLUSIONS:** The FIM-SR scales appear to be reliable and valid measures of independence in 6 specific (self-care, sphincter control, mobility, locomotion, communication, social cognition), and 3 global (motor, cognition, total) areas of functioning in persons with NMD.


**Citation #2:** Who will need long-term care?
Creation and validation of an instrument that identifies older people at risk.

S. Goodlin, C. Boult, T. Bubolz, and L. Chiang

The aim of this study was to create and measure the predictive validity of a screening instrument that identifies older people who are at risk for developing a need for long-term care within a year. This was an observational study, with participants allocated to either a derivation cohort or a validation cohort, in the United States. A nationally representative sample of older community-dwelling Medicare beneficiaries (n = 6,538) participated in the Medicare Current Beneficiary Survey. Questions addressed socio-demographic, functional, health-related, and utilization characteristics in 1991 and 1992, linked to records of Medicare payments for health services during 1991-1992. In the derivation cohort, 14 self-reported characteristics were significant predictors of developing a need for long-term care within 1 year. In the validation cohort, these 14 characteristics identified a high-risk subgroup (18%) that, during the following year, developed a need for long-term care at six times the rate of the low-risk majority. This brief survey instrument identifies a high-risk minority of older people that will, during the following year, develop a need for long-term care at six times the rate of the low-risk majority. This instrument may be useful for targeting at-risk subgroups of older...
populations to receive interventions designed to preserve functional independence and avert the need for long-term care.


Citation #3: Validity of height and weight self-report in Mexican adults: results from the national health and aging study.
J.A. Avila-Funes, L.M. Gutierrez-Robledo, and S. Ponce de Leon-Rosales

BACKGROUND: An adequate nutritional status is essential for maintaining the independence in the elderly. The height and weight self-report is considered a useful alternative for the estimation of body mass index (BMI). The validity of the self-report is an issue that has not been dealt with in developing countries. AIM: To assess the validity of the height and weight self-report in adults Mexican citizens. DESIGN: Transversal study. STUDY POPULATION: 1,707 persons (836 males, 871 females) were asked for their height and weight and were measured. RESULTS: Mean (+/- standard deviation) age was 59.09 +/- 9.86 SD; mean years of education was 5.51 +/- 4.67 years. We found a high correlation between self-reported and measured weight (R^2= 0.837); the difference between both values rose along with the age (from 0.4 kg to 1.74 kg). A systematic difference between self-reported and measured height was found. Self-reported height was over-estimated, and the bias increased along with the age of the subjects (from 1.57 cm to 2.57 cm); further, over-estimation was larger in female individuals (+2.22 cm in female vs. +1.21 cm in male subjects). We calculated a linear model that predicts real height from self-reported height with moderate, although statistically significant results (R^2 = 0.39 y 0.50, for female and male, respectively, p < 0.0001). Knee height was also used to
estimate "adulthood height" and thus, BMI. This method showed age-related dissimilarities, and the linear regression model yielded an unacceptably low correlation (R2 = <0.10). The best method to estimate real BMI was to consider self-reported parameters.

**DISCUSSION:** Height and weight estimation using self-reported parameters is an acceptable method. Its precision is not so high in subjects > 75 years. Height over-estimation is an expected finding congruent with age-related corporal changes. The estimation of height using knee height is not a useful method. **CONCLUSION:** Height and weight self-report is a valid method that may be used to accurately estimate height and weight in Mexican people.


**Citation #4:** Performance of two self-report measures for evaluating obesity and weight loss.

D.L. Patrick, D.M. Bushnell, and M. Rothman

**OBJECTIVES:** To evaluate performance of the Obesity and Weight-Loss Quality-of-Life (OWLQOL) and Weight-Related Symptoms (WRSM) measures. **RESEARCH METHODS AND PROCEDURES:** Four studies of obese persons 18 to 75 years of age were analyzed: a 12-week initial validation study, a clinical trial using blinded endpoint data at 50 to 83 weeks, and community studies conducted in the United States and Europe. Fifty-six initial validation study subjects visited 1 week after screening to evaluate reproducibility. **RESULTS:** Overall, 6107 obese persons completed one assessment, 291 completed follow-up at 12 weeks, and 642 at >50 weeks. Psychometric analyses resulted in a 17-item OWLQOL with a single score tested on five samples that was internally consistent (alpha values > 0.90) and reproducible (intraclass correlation coefficient > 0.95). The OWLQOL score
(higher is better) was associated, as expected, with the symptom measure (lower is better, -0.54), generic quality of life measure (0.53), and measures of physical (0.40) and mental functioning (0.47). The 20-item WRSM was internally consistent (alpha = 0.87) and reproducible (intraclass correlation coefficient = 0.83). The OWLQOL discriminated between genders (p < 0.001), presence of disability days (p < 0.05), levels of BMI (p < 0.05), and levels of symptom bothersomeness (p < 0.001). Evaluation at 12 weeks yielded an effect size for > or = 2.5% weight loss of 0.77 for the OWLQOL and -0.54 for the WRSM. At > or = 50 weeks for > or = 10% weight change, effect sizes were 1.63 and -0.73, respectively. **DISCUSSION:** The OWLQOL and WRSM are brief, valid, reproducible, and responsive self-reported outcomes for evaluating obesity and weight loss.

**Citation #5: Validity of women's self-reports of cancer screening test utilization in a managed care population.**

L.S. Caplan, D.V. McQueen, J.R. Qualters, M. Leff, C. Garrett, and N. Calonge

This study was undertaken to examine the validity of self-reported data on breast and cervical cancer screening behavior. An abbreviated version of the Behavioral Risk Factor Surveillance System telephone survey, including questions on mammography, clinical breast examination (CBE), and Papanicolaou test utilization, was administered to a sample of 480 women aged 40-74 years, enrolled in Kaiser Permanente Colorado for at least 5 years. Screening information reported in the telephone interview was compared with that abstracted from respondents' medical records. The vast majority of women had a mammogram, CBE, and Pap test
according to both self-report and medical record. Sensitivity for determining whether her last test was within 2 years (3 years for Pap test) exceeded 95% for all, whereas specificities were <55%. The percentage of overall agreement between self-reported and recorded information was 88.4% (kappa = 0.62) for mammography, 87.9% (kappa = 0.45) for CBE, and 87.2% (kappa = 0.54) for Pap test. Pearson correlations between self-reported and recorded information for specific time interval since most recent mammogram, CBE, and Pap test were 0.72, 0.58, and 0.65, respectively. Correlation increased greatly when time interval was allowed to vary by +/-1 year. In most cases of disagreement, the self-report underestimated the time since last screening. These results suggest that self-reporting of breast and cervical cancer screening is fairly accurate in this managed care population, although women tend to underestimate the time since their last screening.


Citation #6: The reliability and validity of the self-reported drinking measures in the Army's Health Risk Appraisal survey.

N.S. Bell, J.O. Williams, L. Senier, S.R. Strowman, and P.J. Amoroso

BACKGROUND: The reliability and validity of self-reported drinking behaviors from the Army Health Risk Appraisal (HRA) survey are unknown. METHODS: We compared demographics and health experiences of those who completed the HRA with those who did not (1991-1998). We also evaluated the reliability and validity of eight HRA alcohol-related items, including the CAGE, weekly drinking quantity, and drinking and driving
measures. We used Cohen's kappa and Pearson's r to assess reliability and convergent validity. To assess criterion (predictive) validity, we used proportional hazards and logistical regression models predicting alcohol-related hospitalizations and alcohol-related separations from the Army, respectively. **RESULTS:** A total of 404,966 soldiers completed an HRA. No particular demographic group seems to be over- or underrepresented. Although few respondents skipped alcohol items, those who did tended to be older and of minority race. The alcohol items demonstrate a reasonable degree of reliability, with Cronbach's alpha = 0.69 and test-retest reliability associations in the 0.75-0.80 range for most items over 2- to 30-day interims between surveys. The alcohol measures showed good criterion-related validity: those consuming more than 21 drinks per week were at 6 times the risk for subsequent alcohol-related hospitalization versus those who abstained from drinking (hazard ratio, 6.36; 95% confidence interval=5.79, 6.99). Those who said their friends worried about their drinking were almost 5 times more likely to be discharged due to alcoholism (risk ratio, 4.9; 95% confidence interval=4.00, 6.04) and 6 times more likely to experience an alcohol-related hospitalization (hazard ratio, 6.24; 95% confidence interval=5.74, 6.77). **CONCLUSIONS:** The Army's HRA alcohol items seem to elicit reliable and valid responses. Because HRAs contain identifiers, alcohol use can be linked with subsequent health and occupational outcomes, making the HRA a useful epidemiological research tool. Associations between perceived peer opinions of drinking and subsequent problems deserve further exploration.

*Alcoholism, Clinical and Experimental Research* 27.5 (May 2003): 826-834.
Citation #7: Reliability and validity of self-report of vitamin and mineral supplement use in the vitamins and lifestyle study.


In the United States, dietary supplements contribute a large proportion of micronutrient intakes. Therefore, it is important to collect accurate information on supplement use for studies of micronutrients and disease risk. This report describes the test-retest reliability and validity of a detailed, self-administered mailed questionnaire on vitamin and mineral supplement use. Participants (n = 220) completed the questionnaire at baseline and 3 months later. During an in-person interview, participants provided spot urine and blood samples, and interviewers transcribed nutrient information from their supplement bottle labels. The questionnaire had very good test-retest reliability for mean supplement intake over the past 10 years, with intra-class correlations ranging from 0.69 for beta-carotene to 0.87 for vitamin E. Pearson's correlation coefficients comparing current supplemental intakes from the questionnaire and interviews/label transcriptions were high, ranging from 0.58 for beta-carotene to 0.82 for chromium; however, for some nutrients, median intakes from the questionnaire were slightly lower than from the interviews. Beta-carotene, vitamin C, and vitamin E (alpha-tocopherol) showed clear linear trends of increasing blood concentrations with higher self-reported supplemental intakes (Pearson's correlation coefficients adjusted for potential confounding factors and diet = 0.31, 0.29, and 0.69, respectively; all p < 0.0001). Creatinine-adjusted spot urinary calcium values were not associated with supplemental calcium intakes (Pearson's r = -0.07). This self-administered questionnaire
demonstrated high reproducibility and validity for collecting detailed information on supplement use.


Citation #8: Validity of self-reported height and weight in 4808 EPIC-Oxford participants.
E.A. Spencer, P.N. Appleby, G.K. Davey, and T.J. Key

OBJECTIVE: To assess the validity of self-reported height and weight by comparison with measured height and weight in a sample of middle-aged men and women, and to determine the extent of misclassification of body mass index (BMI) arising from differences between self-reported and measured values. DESIGN: Analysis of self-reported and measured height and weight data from participants in the Oxford cohort of the European Prospective Investigation into Cancer and Nutrition (EPIC-Oxford). SUBJECTS: Four thousand eight hundred and eight British men and women aged 35-76 years. RESULTS: Spearman rank correlations between self-reported and measured height, weight and BMI were high (r > 0.9, P < 0.0001). Height was overestimated by a mean of 1.23 (95% confidence interval (CI) 1.11-1.34) cm in men and 0.60 (0.51-0.70) cm in women; the extent of overestimation was greater in older men and women, shorter men and heavier women. Weight was underestimated by a mean of 1.85 (1.72-1.99) kg in men and 1.40 (1.31-1.49) kg in women; the extent of underestimation was greater in heavier men and women, but did not vary with age or height. Using standard categories of BMI, 22.4% of men and 18.0% of women were classified incorrectly based on self-reported height and weight. After correcting the self-reported values using predictive equations derived from a 10% sample of subjects, misclassification decreased.
to 15.2% in men and 13.8% in women. **CONCLUSIONS:** Self-reported height and weight data are valid for identifying relationships in epidemiological studies. In analyses where anthropometric factors are the primary variables of interest, measurements in a representative sample of the study population can be used to improve the accuracy of estimates of height, weight and BMI.


**Citation #9: Self-report of high cholesterol: determinants of validity in U.S. adults.**

S. Natarajan, S.R. Lipsitz, and P.J. Nietert

**BACKGROUND:** Hypercholesterolemia is a major cardiovascular risk factor, and cholesterol awareness is important in both clinical practice and in public health. We evaluated the validity of self-reported hypercholesterolemia and identified determinants of validity. **METHODS:** The study design was a cross-sectional survey, from 1988 to 1994, of adult participants (N=8236) from the Third National Health and Nutrition Examination Survey for whom self-report of hypercholesterolemia and serum measurement were available. Sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for self-reported hypercholesterolemia were calculated using total cholesterol $\geq 5.17$ mmol/L (200 mg/dL) and/or taking cholesterol-lowering medication as the criterion standard. **RESULTS:** Overall test characteristics for self-report were sensitivity, 51%; specificity, 89%; PPV, 87%; and NPV, 55%. Sensitivity of self-report was higher among older subjects and non-Hispanic whites, specificity was higher among subjects with $>12$ years of education, PPV was higher in older subjects, and NPV was higher in younger subjects and in those with $>12$ years of education. Using higher cholesterol
thresholds to define hypercholesterolemia led to higher sensitivity, lower specificity, lower PPV, and higher NPV. Sociodemographic and anthropometric predictors of validity were identified by logistic regression. **CONCLUSIONS:** Due to low sensitivity, self-reported hypercholesterolemia should be used with caution, both during the patient encounter and for surveillance of trends in hypercholesterolemia in the absence of measured cholesterol levels. Specificity is consistently much higher than sensitivity. The high PPV may be of use in certain clinical situations. Such validation studies should form the foundation for future research based on self-report.


**Citation #10:** Effects of age on validity of self-reported height, weight, and body mass index: findings from the Third National Health and Nutrition Examination Survey, 1988-1994.

M.F. Kuczmarski, R.J. Kuczmarski, and M. Najjar

**OBJECTIVE:** To compare self-reported to measured heights and weights of adults examined in the Third National Health and Nutrition Examination Survey (NHANES III), and to determine to what extent body mass index (BMI) calculated from self-reported heights and weights affects estimates of overweight prevalence compared with BMI calculated from measured values. **DESIGN:** A complex sample design was used in NHANES III to obtain a nationally representative sample of the U.S. civilian, noninstitutionalized population. During household interviews, survey respondents were asked their height and weight. Trained health technicians subsequently measured height and weight using standardized procedures and equipment. **SUBJECTS:** The
analytical sample consisted of 7,772 men and 8,801 women 20 years old and older. **STATISTICAL ANALYSES**

**PERFORMED:** Only persons with measured and self-reported heights and weights were included in the analysis, and statistical sampling weights were applied. t Tests, Pearson product moment correlation coefficients, sensitivity, and specificity analyses were used to determine the validity of self-reported measurements and prevalence estimates of overweight, defined as BMI of 25 or greater. **RESULTS:** Age is an important factor in classifying weight, height, BMI, and overweight from self-reports. Statistically significant differences were found for the mean error (measured-self-reported values) for height and BMI that were notably larger for older age groups. For example, the mean error for height ranged from 2.92 to 4.50 cm for women and from 3.06 to 4.29 cm for men, 70 years and older. Despite the high correlation between measured and self-reported data, the prevalence of overweight calculated from measured values was higher than that calculated from self-reported values among older adults. When calculated with self-reported height, BMI was one unit lower than when calculated from measured height for persons \( \geq 70 \) years. Specificity was high but sensitivity decreased with increasing age cohorts. Regression equations are provided to determine actual height from self-reported values for older adults.

**CONCLUSION/APPLICATIONS:** Self-reported heights and weights can be used with younger adults, but they have limitations for older adults, ages \( \geq 60 \) years. In research studies and in clinical settings involving older adults, failure to measure height and weight can result in subsequent misclassification of overweight status. Therefore, registered dietitians are encouraged to obtain a measured weight and height using a calibrated scale and stadiometer.
Citation #11: How accurately are height, weight and leg length reported by the elderly, and how closely are they related to measurements recorded in childhood?
D. Gunnell, L. Berney, P. Holland, M. Maynard, D. Blane, S. Frankel, and G.D. Smith

BACKGROUND: This paper examines (1) the accuracy of self-reported height, leg length and weight in a group of subjects aged 56-78; (2) whether recent measurement of height and weight influences the accuracy of self-reporting and (3) associations between childhood and adult height, leg length and BMI measured in old age. METHODS: All 3182 surviving members of the Boyd Orr cohort were sent postal questionnaires in 1997-1998 and a sub-sample (294) was also clinically examined. RESULTS: Self-reported height was overestimated and body mass index (BMI), based on reported height and weight, underestimated. The mean difference between self-reported and measured values was for height: 2.1 cm in males and 1.7 cm in females; for BMI the difference was -1.3 kg/m(2) in males and -1.2 kg/m(2) in females. Shorter individuals and older subjects tended to over-report their height more than others. The overweight under-reported their weight to a greater extent. Recent measurement appeared to decrease over-reporting of height but not weight. Correlations between self-report and measured height and BMI were generally over 0.90, but weaker for leg length (r = 0.70 in males and 0.71 in females). Adult height and leg length were quite closely related to their relative values in childhood (correlation coefficients ranged from 0.66 to 0.84), but associations between adult and childhood
BMI were weak (r = 0.19 in males and 0.21 in females).

CONCLUSIONS: Self-reported measures of height and weight may be used in studies of the elderly although systematic reporting errors may bias effect estimates. As overweight individuals tend to under-report and the short and underweight tend to over-report, studies investigating associations of disease with height and weight using self-reported measures will underestimate effects. The weak associations between childhood and adult BMI indicate that associations between childhood adiposity and adult cardiovascular disease found in this cohort may reflect the specific effect of childhood overweight, rather than its persistence into adulthood. This suggests that avoidance of adiposity may be as important in childhood as in adulthood.


Citation #12: Validation of self-reported chronic conditions and health services in a managed care population.


BACKGROUND: Self-reported data are commonly used to estimate the prevalence of health conditions and the use of preventive health services in the population, but the validity of such data is often questioned. METHODS: The Behavioral Risk Factor Survey (BRFS) was administered by telephone to a stratified, random sample of health maintenance organization (HMO) subscribers in Colorado in 1993, and self-reports were compared with HMO medical records for 599 adults aged >21. Sensitivity and specificity were calculated for three chronic conditions and use of six preventive services. RESULTS: Sensitivity was highest for hypertension (83%), moderate for diabetes (73%), and lowest for hypercholesterolemia (59%);
specificity was >80% for all three conditions. Sensitivity ranged from 86% to 99% for influenza immunization, clinical breast examination, blood cholesterol screening, mammography, Pap test, and blood pressure screening; specificity was <75% for all preventive services. **CONCLUSIONS:** Self-reports are reasonably accurate for certain chronic conditions and for routine screening exams and can provide a useful estimate for broad measures of population prevalence.


**Citation #13:** Differences in morbidity measures and risk factor identification using multiple data sources: the case of stroke.

J. Madans, C. Reuben, S. Rothwell, M. Eberhardt, and J. Feldman

**BACKGROUND:** Epidemiologic studies utilize medical information from a variety of sources. These include subject or proxy interviews, medical records, death certificates and administrative records. Since the choice of data source may affect the validity of study results, it is important to understand the effect of different case-ascertainment methodologies on estimates of risk.

**METHODS:** The NHANES I Epidemiologic Follow-up Study (NHEFS) contains several sources of information that can be used to define case status. In this report we investigate whether the use of seven different algorithms for case ascertainment, each based on different combinations of data sources, results in substantive differences in the estimates of incidence rates and relative risks associated with selected, documented, risk factors for stroke.

**RESULTS:** The seven different models of case identification gave very different estimates of stroke incidence. However, the
characteristics of the cases defined by the models, except for cases identified by death certificate only, were remarkably similar. There was also remarkable similarity in relative risks obtained from six of the seven models. The model using only death certificate information generally produced higher relative risk estimates.

**CONCLUSIONS:** Despite wide variations in the estimates of incidence, characteristics of the cases using different case definition were remarkably similar, as were the risks associated with stroke incidence. The main difference occurred when cases were identified from the death certificate only. These results furnish some evidence that analyses based on self report can provide valid, useful information.


**Citation #14: Psychometric evaluation of daytime sleepiness and nocturnal sleep onset scales in a representative community sample.**

E.O. Johnson, N. Breslau, T. Roth, T. Roehrs, and L. Rosenthal

**BACKGROUND:** The public health importance of daytime sleepiness as a risk factor for accidents, interpersonal problems, and decreased productivity has been recognized. However, epidemiologic research on this topic has been limited by the reliance on laboratory measures (i.e., the Multiple Sleep Latency Test-MSLT). Two scales, daytime sleepiness and nocturnal sleep onset, have been identified from the self-report Sleep-Wake Activity Inventory (SWAI) in a clinic sample and validated against the MSLT. This study evaluates the replicability of the two scales in a population sample and assesses potential thresholds in scale scores that distinguish normal from pathologic levels of daytime sleepiness and difficulty falling asleep. **METHODS:** The sample consisted of 2181 subjects 18-45 years old in the Detroit
metropolitan area. All sleep characteristic information covered the 2 weeks prior to interview. Split-half sample factor analyses were conducted to assess replicability of the results. Distribution of scale scores and their relation to construct validity variables were used to evaluate possible thresholds. **RESULTS:** A two-factor model appeared to best account for the variation among the 12 items from the SWAI. The two factors accounted for 50% of the variance in both split-half sample analyses. The revised eight-item daytime sleepiness and two-item nocturnal sleep onset scales showed good and fair internal consistency respectively across both split-half samples. There appeared to be a "natural break" in daytime sleepiness scale scores that was associated with a substantial and consistent change in number of hours slept. No breaks appeared in nocturnal sleep onset scores. **CONCLUSIONS:** This study replicated the results of the clinic-based study and suggested a potentially useful diagnostic threshold for self-report excessive daytime sleepiness. Epidemiology of sleep depends on the ability to move from the laboratory to population surveys in reliable and valid ways. Development of self-report is a step in that direction.


**Citation #15: Validity of a self-reported history of doctor-diagnosed angina.**


The objective of this study was to assess the validity of a self-reported history of doctor-diagnosed angina in population-based studies in men. Subjects were 5789 men from the British Regional Heart Study who reported being without an angina diagnosis at entry (1978-1980) and were alive at the end of 1992, aged 52 to 75.
years. In 1992, subjects were asked in a self-administered questionnaire if they recalled ever having had a doctor diagnosis of angina. Self-report of diagnosed angina was compared with general practice (GP) record of angina obtained from reviews of medical records from study entry to the end of 1992. Men were followed for a further 3 years from 1992 for major ischemic heart disease events. The prevalence of diagnosed angina in 1992 was 10.1% according to self-reported history and 8.9% according to GP record review. There was substantial agreement between the two sources of information: 80% of men with a GP record of angina reported their diagnosis, and 70% of men who reported an angina diagnosis had confirmation of this from the record review. When all ischemic heart disease (angina or myocardial infarction) was considered, agreement was higher. Genuine angina was likely in many of the 177 men who had self-reported angina not confirmed by the GP record review: 78 had an ischemic heart disease history (myocardial infarction or coronary revascularization) identified by the review, and 31 had a GP record of angina after 1992. Angina symptoms, nitrate use, cardiological investigation, and surgical intervention for angina compared between agreement groups showed a very consistent pattern. All these indicators of angina were most common in men with both self-report and GP record of angina, least common in men with neither self-report nor GP record of angina, but had a substantially higher prevalence in men with self-reported angina only than in those with GP-recorded angina only. After 3 years follow-up from 1992, 9.5% of men with both self-report and GP record of angina, and 11.3% of men with self-reported angina only had experienced a new major ischemic heart disease event; compared to 5.7% of men with a GP record of angina only and 2.7% of those without angina by either criteria. This pattern of risk remained similar after adjustment for age and
previous myocardial infarction. **These results suggest that self-reported history of a doctor diagnosis of angina is a valid measure of diagnosed angina in population-based studies in men.**


Citation #16: Accuracy of self-report of mammography and Pap smear in a low-income urban population.

P.G. McGovern, N. Lurie, K.L. Margolis, and J.S. Slater

**BACKGROUND:** Cancer screening history can often be obtained only by self-report, particularly for disadvantaged populations. We examined the accuracy of self-report of mammography and Pap smear for an urban, low-income population. **METHODS:** Women attending non-primary care clinics (mostly surgery and orthopedics) at a large public teaching hospital in Minneapolis between July 1992 and May 1993 were queried about their screening history (n = 477). The women were interviewed by a trained peer-recruiter and asked whether they had ever heard of a Pap smear or mammogram, whether they had ever had one, where it was done, and when the last one was. We verified self-report by checking medical records where the test was performed. **RESULTS:** The positive and negative predictive value of recall of mammography in the previous year was 72.4% and 90.6%, respectively. The figures for Pap smear recall were somewhat lower, 65.5% and 85.9%, respectively. We found a record of a mammogram in 88% of women able to recall the year. Of these, slightly over two-thirds recalled their mammogram in the same year as their record indicated. Inaccurate recalls were more commonly of the "telescoping" type, i.e., tests were recalled as having occurred more recently than was the case. Recall was substantially better for recent tests. Results for Pap smear recalls
were broadly similar. **CONCLUSIONS:** The accuracy of self-report of mammography and Pap smear is relatively poor for medical practice but is acceptable in population surveys with appropriate correction for over-reporting.


**Citation #17: Validity of self-reported hypertension in the National Health and Nutrition Examination Survey III, 1988-1991.**

C.M. Vargas, V.L. Burt, R.F. Gillum, and E.R. Pamuk

**BACKGROUND:** The National Health and Nutrition Examination Survey (NHANES) is the main data source for hypertension surveillance. However, because of a gap of almost 10 years between each NHANES, self-reported data from annual surveys need to be examined as an alternative data source. This study analyzes the validity of self-reported hypertension in a national sample of non-Hispanic whites, non-Hispanic blacks, and Mexican-Americans. **METHODS:** Sensitivity, specificity, and predictive values positive (PVP) and negative (PVN) of self-reported hypertension were calculated against two definitions of hypertension: the definition recommended by the Third Joint National Committee on Hypertension, JNC III (blood pressure > or = 140/90 and/or taking antihypertension medication) and a broader definition including control with lifestyle modifications. Data used come from the NHANES III, 1988-1991. **RESULTS:** Overall test characteristics using the JNC III definition are sensitivity 71%, specificity 90%, PVP 72%, and PVN 89%. Test characteristics were consistently higher for the broad than for the JNC III definition. Validity of self-reported hypertension is higher among women than among men and among persons with a medical visit
during the past year than among those with no visits: validity was lowest among Mexican-American men. Due to the similarity between sensitivity and PVP, the prevalence of self-reported hypertension is nearly equal to the prevalence of JNC III-defined hypertension. **CONCLUSIONS:** Self-reported hypertension may be used for surveillance of hypertension trends, in the absence of measured blood pressure, among non-Hispanic whites and non-Hispanic black women and persons with a medical visit in the past year. Validation should be repeated with each NHANES.


**Citation #18: Estimating the burden of disease. Comparing administrative data and self-reports.**

J.R. Robinson, T.K. Young, L.L. Roos, and D.E. Gelskey

**OBJECTIVES:** A cardiovascular health survey of a representative sample of the adult population of Manitoba, Canada was combined with the provincial health insurance claims database to determine the accuracy of survey questions in detecting cases of diabetes, hypertension, ischemic heart disease, stroke, and hypercholesterolemia. **METHODS:** Of 2,792 subjects in the survey, 97.7% were linked successfully using a scrambled personal health insurance number. Hospital and physician claims were extracted for these individuals for the 3-year period before the survey. **RESULTS:** The authors found no benefits to using restrictive criteria for entrance into the study (i.e., requiring more than one diagnosis to define a case). Using additional years of data increased agreement between data sources. Kappa values indicated high levels of agreement between administrative data and self-reports for diabetes (0.72) and hypertension (0.59); kappa values were approximately 0.4 for the other conditions. Using administrative data as the "gold standard," specificity was
generally very high, although cases with hypertension and hypercholesterolemia (diagnosed primarily by laboratory or physical measurement) were associated with a lower specificity than the other conditions. Sensitivity varied markedly and was lowest for "other heart disease" and "stroke". For diabetes and hypertension, inclusion criteria calling for more than one diagnosis reduced the accuracy of case identification, whereas increasing the number of years of data increased accuracy of identification. For diabetes and hypertension, self-reports were fairly accurate in detecting "true" past history of the illness based on physician diagnosis recorded on insurance claims. CONCLUSIONS: This study demonstrates the feasibility of linking a large health survey with administrative data and the validity of self-reports in estimating the prevalence of chronic diseases, especially diabetes and hypertension. A linked data set offers unusual opportunities for epidemiologic and health services research in a defined population.

Medical Care 35.9 (Sept. 1997): 932-947.

Citation #19: Validity and reliability of self-reported drinking behavior: dealing with the problem of response bias.
B.G. Embree, and P.C. Whitehead

This work assesses the validity and reliability of self-reported survey data on drinking behavior. There is evidence to suggest that data are adversely affected by bias from underreporting. This bias affects the validity of measures of consumption of alcohol and can have deleterious effects on the results of some forms of statistical estimation. Data for this study were collected at an isolated military base. The remoteness of this site and the fact that it is a military station made it possible to estimate the actual level of consumption of alcohol for the population by assessing apparent consumption
through officially recorded sales of alcohol. The results of eight measures of consumption of alcohol were compared with apparent consumption, as established by documented sales, and the validity and reliability of the various measures were determined using the classical correlational approach. The validity and reliability of the data generated by the self-report survey were also analyzed using LISREL, the measurement model in particular. The results indicate that various instruments used to assess the consumption of alcohol produce very different outcomes in terms of their validity and reliability, some questions being considerably more valid and reliable than others. Two of the more salient characteristics of questions that affect validity and reliability were isolated, namely a question's ability to aid recall and its ability to mitigate the effects of persons providing socially desirable responses. The LISREL results show that these are two underlying factors for the measurement of the consumption of alcohol. It is concluded that questions that produce valid and reliable responses do so for identifiable reasons, and measurement instruments can be improved by incorporating particular features.

F. Glossary of Terms

The studies highlighted in this document contain a wide variety of research and evaluation terms. This glossary is intended to assist the reader with understanding the key terms of this *White Paper*.

**ANOVA:** A major statistical technique for analysis of variance that is used in the inferential analysis of data.

**Attrition:** The loss of sample members from the experiment during the time period of the study from one of many different causes.

**Blocks:** Groups of units of subjects that are as homogeneous within a group as possible. Blocks are frequently compared with each other to determine similarities and differences.

**Chi-Square:** An inferential statistical test used with nominal and ordinal level data that is a non-parametric statistic. This is a very common test of attribution or causality between events or actions and observed outcomes.

**Comparative Evaluation:** The type of evaluation that relies on the use of normative reference points in the development of the interpretation of the data collected in the course of the evaluation.

**Concomitant Variables:** Those variables that are observable or measurable phenomena that are statistically associated with the responses of principal or major interest. For example, the cholesterol level of someone three months after they have been tested and counseled for cholesterol reduction.
Construct Validity: The meaningfulness of the attribute or variable being examined or under investigation. The extent to which the methods of observation actually reflects it and the validity of any resulting interpretation. The significance lies in the valid linkage of a more global concept, such as “well-being” with something as specific as the measurement of blood pressure.

Content Validity: The extent to which the test or measurement method elicits a reasonable sample of the totality of behaviors that characterize the variable of interest. For example, is VO² uptake a valid singular measure of cardiopulmonary function?

Criterion Validity: The extent to which measurements and observations associated with particular methods of measurement are consonant with a generally accepted and presumably more accurate standard. For example, how does percent body fat measures using electrical impedance compare with hydrostatic weighing results?

Dependent Variable: The variable that is modified or manipulated in an experimental situation. For example a smoking cessation course is the dependent variable in evaluating changes in smoking behavior participants and non-participants.

Differential Attrition Effects: A threat to internal validity that results from differential loss rates for experimental subjects versus control groups producing differences in mean scores and other parameters.
**Evaluation:** Evaluation is the determination of the relative value of a phenomenon by appraisal or comparison with a standard. To evaluate is to obtain sound information upon which to base decisions. Program evaluation is the effort to evaluate the efficiency, effectiveness and implications of a set of primarily organized or linked activities.

**Evaluation Research:** The process of applying scientific methods to accumulate reliable and valid information regarding the effect of an intervention on selected variables within a defined population.

**Experimental Design:** The plan or structure for conducting an experiment. It includes such things as sample size, selection method, type and level of data to be collected, statistical tests to be applied, and how the results will be used.

**External Validity:** The validity of the conclusions and inferences applied from the group of subjects studied to another group of subjects not studied or evaluated. External validity answers the question of whether the results can be applied to another group with significant validity.

**Factors:** The variables under examination.

**Field Testing:** The testing of an evaluation design prior to instituting the full system. For example, using a wellness program evaluation survey with a small number of volunteers to examine whether the questions are understandable and valid. This process is intended to reduce the “bugs” in the instrument or with the methodology used.
Formative Evaluation
Evaluation that is performed to monitor and help refine a particular set of activities. Usually there is no identified norm that is used as a comparison point for the activity under examination.

Gamma:
A descriptive statistic used with ordinal type data to describe the strength of association.

History:
A threat to internal validity where events other than the experimental treatment occur between pretest and post-test and provide alternative explanations of observed effects.

Homogeneity:
The extent to which all the subjects in an identified group are consistent or alike. A measure of similarity.

Heterogeneity:
The extent to which all the subjects in an identified group are not consistent or alike. A measure of dissimilarity.

Hypothesis:
The statement of relationship that is being tested.

Independent Variable:
The type of variable that is consistently present in both experimental and control groups.

Instability:
A threat to internal validity that results from the unreliability of measures, fluctuations in sampling methodology or autonomous instability in measure elements. This is the only threat to which statistical tests of significance are relevant.

Instrumentation:
A threat to internal validity that results from changes in the calibration of a measuring instrument or changes in observers or scores producing changes in measurements.
Internal Validity: The validity of results, findings and conclusions drawn from evaluation of a specific group of subjects or programs. Internal validity answers the question of whether the results that were obtained from the evaluation were in fact valid for the subject population under study.

Interval: The type of data that has a numerical measure difference between items.

Irrelevance of Treatment Effects: A threat to external validity resulting from complex experimental interrelationships of treatment interventions that may fail to include those components that are actually responsible for the effects. For example the instructor effect may have a lot to do with behavioral adherence but it may not be measured or factored into the evaluation design.

Irrelevant Responsiveness of Measures: A threat to external validity that results from complex measures that contain irrelevant components that may produce apparent effects.

Lambda: A statistical test used with nominal type data that measures the predictability of association.

Maturation: A threat to internal validity that reflects processes within the subjects that accounts for the observed effects rather than the treatment intervention.

Measures of Central Tendency: The types of measurements that reflect the distribution of data among measurable variables. The most frequently used measures are mean, median and mode. Mean is the arithmetic average of the score involved, median is the middle score among a group of scores and mode is the most frequently occurring score or value.
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple-Treatment Effects:</strong></td>
<td>A threat to external validity that results from multiple treatments that are jointly applied and produce changes that are not directly attributable to the treatments when applied in a singular fashion.</td>
</tr>
<tr>
<td><strong>Monitoring:</strong></td>
<td>To monitor is to watch, observe or check for a specific purpose. It is usually a systematic process that collects information over time.</td>
</tr>
<tr>
<td><strong>Nominal:</strong></td>
<td>The lowest level of data that describes an either-or situation, such as male-female or smoker-non-smoker.</td>
</tr>
<tr>
<td><strong>Non-parametric Statistic:</strong></td>
<td>A type of statistics without a normal distribution. A skewed distribution is assumed.</td>
</tr>
<tr>
<td><strong>Ordinal:</strong></td>
<td>The type of data which has a rank order assignment. For example, low, medium, high.</td>
</tr>
<tr>
<td><strong>Parametric:</strong></td>
<td>A type of statistics with a normal distribution assumed.</td>
</tr>
<tr>
<td><strong>Phi:</strong></td>
<td>A descriptive statistic test used with nominal or ordinal type data.</td>
</tr>
<tr>
<td><strong>Program:</strong></td>
<td>The combination of one or more projects intended to achieve, in a goal oriented manner, some systematic change.</td>
</tr>
<tr>
<td><strong>Project:</strong></td>
<td>The implementation and organization of actions that are directed to achieve a defined set of objectives.</td>
</tr>
<tr>
<td><strong>Randomization:</strong></td>
<td>The provision of equal probability that any single case or subject could be assigned to the experimental or control group.</td>
</tr>
</tbody>
</table>
Regression: A threat to internal validity that results from pseudo-shifts which occurs because subjects have been selected based on their extreme scores. The phenomenon is also labeled as regression to the mean.

Reliability: The phenomenon that involves the assumption that any observation contains a systematic and a random component. The degree of reliability is the degree to which the test or measurement accurately reflects what it intends to measure.

Response Variables: The characteristics of the sample members that will be affected by the experimental intervention.

Selection: A threat to internal validity that results from biases incidental to selective recruiting of subjects or differential recruitment of comparison groups producing artificial differences in mean scores and other parameters.

Selection-Maturation Interaction: A threat to internal validity that reflects selection biases resulting in differential rates of maturation or autonomous change.

Selection-Treatment Interaction Effects: A threat to external validity that results from an unrepresentative responsiveness among the experimental or treatment group.

Significance: The probability that a hypothesis is correct.

Situation Effects: A threat to external validity that results from an artificial series of circumstances not innately related to the subject population. An example would be the “Hawthorne effect”, or the effect attributed to the process of being monitored.

Summative Evaluation: The performance of evaluation so that it assesses the overall impact of the activity being studied.
Target Population: The population or subjects at interest to the activity being evaluated.

Testing Interaction: A threat to external validity that results from use of a pretest that materially alters the sensitivity of the subjects or their responsiveness to the experimental intervention thus making any comparison with populations that are not pretested.

Testing-Treatment Interaction Invalidity: A threat to external validity that results from interaction between testing and the treatment.

Threats to Validity: The specific types of problems that weaken internal and external validity of evaluation results.

Validity: Whether the results of the evaluation really measure what is purported to have been evaluated. The two types of validity are internal validity which deals with statements made about the subjects under study and external validity that deals with the application of findings to groups or situations outside those that were studied.

Variables: The characteristics of the subjects that are examined in a formal evaluation. The two major types are independent and dependent variables.
G. Bibliography on the Validity and Reliability of HRAs


34. Kuczmarski, M.F., R.J. Kuczmarski, and M. Najjar. “Effects of age on validity of self-reported height, weight, and body mass index: findings from the Third National


59. Wang, Feifei, Timothy McDonald, Laura Champagne, and Dee W. Edington. “Relationship Between BMI, Physical


