Validation of Colorectal Cancer Screening Behaviors

Jane G. Zapka
Medical University of South Carolina, Charleston, South Carolina

The importance of identifying efficacious and effective strategies to improve the use of evidence-based cancer screening tests and reduce disparities is well documented (1). Among many methodologic challenges to rigorous study design (2) are several related to obtaining reliable and valid measures of screening behaviors. Meissner and colleagues (3) have made several major recommendations for future research studies that promote the use of cancer screening. Included in these recommendations are the importance of developing and testing standardized behavioral outcome measures, and the need for additional studies of reliability and validity in different population subgroups.

Vernon and colleagues (2) have emphasized that selection and measurement of primary outcome measures is critical to the credibility of studies of systems and behavioral change interventions. The authors identified seven important “lessons learned” for cancer screening research. Among these lessons were (a) the need for agreement regarding conceptual and operational definitions of behavioral outcomes; and (b) that studies, using self-reported cancer-screening participation, should assess reliability and validity and quantify measurement error and bias in a broad range of respondents.

With regard to the first point, the need for agreement on operational definitions, progress has been made with respect to colorectal cancer (CRC) screening behaviors. Convened by the National Cancer Institute with the support and participation of the American Cancer Society and the Centers for Disease Control and Prevention, a group of experts developed a uniform description of tests and items to measure CRC screening behaviors. An instrument was subjected to cognitive pretesting and survey items for the four CRC screening tests were recommended (4). The recommendations for further research, the lessons learned, and the recommended survey items provide the backdrop for screening studies reported in this issue of Cancer Epidemiology Biomarkers & Prevention.

This issue contains a group of thoughtful articles highlighting the challenges to obtaining valid and generalizable measures of CRC behaviors to document trends and gaps in maximizing screening rates among diverse populations. The authors collectively make a convincing case for the importance of research related to self-report measures. The Health Insurance Portability and Accountability Act of 1996 has increased the need for, and subsequent use of, self reports because of the limitations on access to medical records (5). Further limiting such access is the fact that individuals may receive across several practices over time in the U.S. health-care system. The CRC screening guidelines pose additional challenges, given the number of test options, several of which are done in specialty practice settings.

Rauscher and colleagues (6) set the stage with a meta-analyses of validation studies of seven screening tests for four cancer types (breast, cervical, prostate, and CRC). The overall findings from this collection of reports are that

- validity of self-report is generally respectable, but may vary by test (7, 8);
- validity may also vary by patient characteristics, such as age, education, and family history (7); by race (6); by intervention assignment (9); and by time (9); and
- although adjustment can be made to national screening prevalence estimates, overreporting (due to telescoping or social desirability) will continue to challenge practitioners’ ability to determine the accuracy of individual patient reports.

Several additional observations from the seven articles in this section are noteworthy.

Survey Mode May or May Not Matter. Several articles consider the mode of data collection and the resulting potential error. Vernon et al. (8) found little variation in validation rates by mode (telephone, mail, or face to face), yet the overall response rate among those contacted and eligible was modest (46%), and the study was conducted in a largely Caucasian, educated population. In a statewide survey (mail and telephone), Beebe and colleagues (10) found in a mailed survey, that asking about intention to be screened, before asking about actual recent screening behavior, statistically significantly reduced the prevalence of claims of recent screening, perhaps due to lowered pressure to give a socially desirable response. Attention to item order with adequate pretesting is needed (11). Standardized operational definitions and items are a strength, but item ordering and context need consideration and additional testing, and may vary by mode.

Measurement Studies Are Needed in Multiple Population Subgroups. The articles consider a number of different populations, including U.S. veterans (7), first-degree relatives (12), the general adult population in five Minnesota counties (9), North Carolina Medicare enrollees (13), a random digit dial sample of Minnesota.
Validation of Colorectal Cancer Screening Behaviors

 adultos (10), and patients of a large Texas clinic system (8).

The findings of Bastani et al. (12) underscore the lesson
that CRC screening behaviors are complex, cognitive
tasks, and the growing cultural diversity of populations
adds to this complexity. As noted in the meta-analyses of
multiple screening tests (6), racial/ethnic variation in
estimates of sensitivity and specificity can mask dispari-
ties in prevalence estimates. This complexity is further
heightened by the opportunistic approach to screening in
the United States, where testing and the process of
screening can take place in numerous settings that lack
integration and communication (12, 14). Further, the
uneven quality and feasibility of record access in
different health care settings poses another challenge to

validation studies. Finally, it is difficult to collect
validation information for groups that are often the
target of intervention, such as samples recruited from
cancer registries or community groups (12).

The “Gold Standard” May Not Be Gold. Schenck et al.
(13) highlight the limits of all three data sources—
Medicare claims, self-report, and medical records. The
meta-analysis of screening studies (6) discusses the
likelihood that the quality of documentation of true
screening history influences estimates of specificity and
positive predictive value. As emphasis on improvement
in information technologies grows (14-16), health system
and practice data may improve and reflect more accurate
screening histories.

Numerous Analytic Methods to Assess Validity Are
Applicable and Similar Reporting Should Be Encour-
aged. The articles in this section make comparisons of
records and reports by computing concordance, sensi-
tivity, specificity, and report-to-records ratio. Addition-
ally, Partin et al. (7) explored the effect of a liberal
versus a strict time interval on concordance, and Jones
et al. (9) investigated forward telescoping. Using liberal
window estimates may be preferable (7) and more
realistic in terms of waiting times for screening; how-
ever, little is known about the effect on testing
efficacy of minor delays. How missing data are handled
is also a potential source of variability across studies.
Validity and reliability may be affected by timing of
report since testing. In the meta-analysis (6), only 29 of
55 studies were included because limited analyses in
some studies precluded comparisons. Just as standard-
ization is advocated for survey items, more standardi-
zation in terms of reporting study results would also
move the field forward.

Researchers and Practitioners Can Consider Oppor-
tunities to Do Experimental Validation Studies within
Other Parent Studies. Methods for validation studies are
difficult, costly, and may not be feasible (12). Schenck
et al. (13) illustrate the considerable resources needed to
find multiple health-care providers in various locations
to provide complete data on a single study participant.
One approach that was taken by several investigators
(9, 10, 12, 13) was to nest validation studies in other
projects.

Overreporting of screening was common, but it is
more problematic when it reflects systematic error such
that, for example, intervention-group participants over-
report to a greater extent than control-group partici-
pants, thus producing spurious results on intervention
efficacy (9). Design features, including interviewer skill,
respondent procedures, call-back/reminder rules, and
refusal conversion all have implications for reducing
error (17). Examples are the use of active refusal
strategy (8), the “total design” strategy for mailed
surveys (8, 18), blinding of auditors (9), and quality
monitoring of auditors (13) and interviewers. Formatting
and appearance can affect decision to participate
(8). Indeed, proper survey development and implement-
ation requires extensive understanding of design and
method details.

High Response Proportions Are Difficult to Achieve and
Failure Represents a Potential Source of Selection
Bias in Validation Studies. Representativeness can be
lost at various stages during the accrual process, for
example, enrollment, follow-up, or permission to access
medical records. Participation may vary by respondent
characteristics, such as race, and marital status (7),
and by screening prevalence (9). Bastani and colleagues
(12) were able to analyze only 13% of their original
sample and found that insurance status and self-
reported screening were independent predictors of
whether potential participants agreed, verbally, to
participate in the validation study. Furthermore, African
American race, low education, and lower reported
screening were independently related to actively mail-
ing consent to access records. These findings underscore
the need for researchers and service providers to
implement rigorous data collection principles. It is
likely, however, that modest response rates in valida-
tion studies may not be as problematic as in prevalence
studies (19, 20).

Transparency of Method, Measures, and Population
Characteristics in Accrual and in Loss to Follow-up Is
Essential in Reporting. Response proportions are of
concern because of their effect on selection bias and
generalizability. Details of sampling and accrual are
essential to measure bias and compare findings across
studies. Superior models of graphic and tabular report-
ing are exemplified in these articles (7, 8, 12). Construc-
tion of measures must also be clear to compare findings.
For example, even when methods are transparent,
they are not necessarily consistent in these studies
(e.g., measuring screening interval with month/day
versus within recommended time interval). Groves and
colleagues (17) have described, in detail, potential quality
issues in survey research, highlighting measurement
issues (validity, measurement error, and processing
error) as well as sampling issues (coverage, sampling,
response effects, nonresponse error, adjustment error).
Each of these issues includes further challenges and
highlights the need for approaching consensus on
operational definitions.

These Articles Show Methods and Principles that
Should Be Considered in Other Behavioral Research.
Epidemiologic, behavioral intervention, and health
services research are challenged by the need to develop
valid, reliable measures of multiple health behaviors
(21-23). The importance of strong validation studies and
the need for transparency in operational definitions
and measurement construction, to ensure understanding
of limitations and to promote comparability should be
stressed.

Cancer Epidemiol Biomarkers Prev 2008;17(4). April 2008
In summary, the work of the National Cancer Institute, Centers for Disease Control and Prevention, American Cancer Society, and numerous investigators across the country have shown foresight and leadership in encouraging research that develops standardized, reliable, and valid measures of CRC screening behaviors. This effort will contribute immensely to the conduct and dissemination of credible efficacy and effectiveness studies. Additionally, the importance of validation studies is underscored by Rauscher et al., who showed that, when prevalence estimates from national surveys were adjusted using random effects summary estimates, population prevalence estimates decreased and race/ethnicity disparities increased (6). Such data are important both to policy and practice decisions. Clearly, many questionnaire design features affect accuracy of self-reports and many have been investigated in this set of important articles. Progress with CRC screening behaviors represents an exemplary model for other areas of behavioral research.

References