Association for Molecular Pathology
Promoting Clinical Practice, Basic Research, and Education in Molecular Pathology
9650 Rockville Pike, Bethesda, Maryland 20814
Tel: 301-634-7939 ▪ Fax: 301-634-7990 ▪ Email: amp@asip.org ▪ *www.amp.org*

May 6, 2009

AMP Comments before the CMS Medicare Evidence Development & Coverage Advisory Committee (MEDCAC) meeting entitled "Screening Genetic Tests."

Presented by Roger D. Klein, MD JD
Member, AMP Economic Affairs Committee

I. Evidence standards for screening tests

In general the evidence relied upon to support implementation of a screening program should meet higher standards than those used to establish coverage for diagnostic or prognostic tests. First, by definition screening tests are performed on asymptomatic individuals. Therefore, all morbidity, mortality, inconvenience or psychological distress resulting from a screening program is iatrogenic. That is, such harms would not have occurred but for the implementation of the screening program. Consequently, the test as applied should have high true positive and low false positive fractions in order to limit harms and attain favorable risk-benefit and cost-benefitbalances. Second, screening is systematically applied to large numbers of asymptomatic people. Therefore, the public health impacts, both positive and negative, may be profound and the repercussions of errors in effectiveness substantial. Finally, screening programs are expensive. Costs include not only the expenses associated with initial screening, but also those attributable to follow up testing and intervention. Resources applied unproductively are diverted from alternative, more beneficial uses.

II. Evidence standards for genetic vs. non-genetic tests

There should in principle be no difference in the evidence thresholds to which genetic tests are held, as compared with the standards applied to other types of screening tests. As an exception, utilization of the subset of genetic tests derived from genome-wide association studies presents unusual challenges as described in the subsequent paragraph. However, it must be emphasized that universally applicable evidence standards cannot properly be developed. Instead, thresholds must be individualized to the specific purpose under consideration. For example different evidence standards will apply for markers used for risk stratification versus those applied for early disease detection. Appropriate standards vary based upon the seriousness of the underlying disease entity, the potential harms engendered by a screening program, and the sub-populations toward which the program is directed. A positive screening test that leads to an invasive procedure has vastly different implications than a result which merely encourages salutary lifestyle changes.

III. Genome-wide association studies

Genome-wide association studies (GWAS) pose particular problems in evidence evaluation, most notably because of a unique propensity to yield false positive results. Importantly, because GWAS analyze large numbers of markers simultaneously, even a test with perfect analytic performance faces a high likelihood of uncovering false associations. Additional sources of error

include phenotypic misclassification of study participants or population stratification, that is segregation by ancestrally related rather than disease associated markers.

Second, although GWAS involve significant technically complexity, they usually are not performed under CLIA-certified conditions. Therefore, analytically perfect results cannot be assumed.

Third, effects are commonly reported as odds ratios or similar measures. However, because an odds ratio can be associated with a range of true positive and false positive fractions (TPF and FPF), strength of association alone is rarely sufficient to establish suitability as a discriminatory marker.

Fourth publication bias toward positive results tends to exaggerate the strength of associations, especially in the early stages of research. Similarly, effect sizes of true associations tend to be inflated when, as is often the case, putative discoveries have met high thresholds of statistical significance in suboptimally powered studies. In other words, overestimated associations are more likely to meet threshold p values, biasing observed associations upward. Finally, substantial deviation of genotype frequencies from Hardy-Weinberg equilibrium can be a clue to possible genotyping errors that manifest as false positive (or negative) associations.

IV. Analytic validity of genetic tests

Most molecular diagnostic methods used in clinical settings offer excellent analytic performance, which is among the reasons they have been rapidly incorporated into clinical laboratory practice. Prior to clinical use of an assay, analytic sensitivity and specificity should be established. Moreover, tests should demonstrate reproducibility and consistency in performance in response to limited changes in preanalytic and analytic variables. Often, the superiority of newer molecular methods renders previous methodologic gold standards obsolete. However, many tests are performed using a common set of methodologies, techniques, and platforms with which there is an extensive general body of experience. Therefore, published data reporting the properties and performance of particular methodologies in a range of contexts is frequently available. In addition, performance information may be available from proficiency testing and interlaboratory comparison programs. For truly novel methods for which there are limited performance histories, collaborative studies using single, large, carefully selected panels of well-characterized samples that are tested and reported blindly under routine conditions, with the results independently analyzed, are desirable.

V. CLIA regulatory requirements, professional accreditation, and licensing

As a general rule, the analytic performance of molecular genetic tests performed in CLIA-certified laboratories is very high. This is among the key reasons Molecular Pathologists and other clinical laboratorians find molecular diagnostic methods attractive. For all clinical tests that are lawfully performed in the United States, the CLIA regulations, and commonly professional accreditation processes such as that of the College of American Pathologists (CAP), combine with professional licensing and certification, to ensure the performance characteristics of diagnostic laboratory tests.

VI. Screening test components

AMP Comments, MEDCAC - May 6, 2009 page 3 of 3

The three major screening components include the disease, the test, and population to be screened. To be the subject of a screening program, a disease must have substantial public health importance, its natural history must be known, and there must be effective interventions that are available and accessible to the screened population. Because of the large numbers of patients to be screened, such tests must offer high analytic and clinical sensitivity and specificity.

Clinical sensitivity, the proportion of early stage or high risk individuals who will have positive screening results, is central to the effectiveness of a test. Clinical specificity, the proportion of disease-free or low risk persons who have negative screening results bears strongly upon the potential harms and unnecessary costs of a screening program. Measures as such positive and negative predictive values incorporate the likelihood of disease or lack thereof in individuals who respectively test positive or negative for a marker. These latter parameters, which depend on the prevalence of the disease in the screened population, more directly bear upon the effectiveness of the screening test.

In addition, two important biases are inextricably linked to the screening process. Patients whose illnesses are detected earlier in the course of their diseases will, in the absence of any useful intervention, display longer survival times than individuals diagnosed later in the course of their diseases. This property is known as lead time bias. Length bias denotes the tendency for patients with less severe disease to be disproportionately represented within a population of asymptomatic patients. For example, patients with slow growing tumors that are as yet unrecognized will appear with relatively greater frequency among screened patients, because those with aggressive tumors are more likely to have progressed to advanced stages during which they will be diagnosed. These two biases can make an ineffective screening program appear to have value. The difficulty in correcting for biases has legitimately caused some to argue that randomized controlled trials should support all new screening programs.

Finally, the ‘critical point’ is the point beyond which treatment does not improve disease outcomes. If the critical point is well beyond the asymptomatic period, and patients who are treated after clinical detection do as well as those who are detected with screening, there is no value to the screening. Conversely, if the critical point extends too early in the asymptomatic period, it will be difficult to identify patients at early enough stages in the course of their diseases to positively affect their outcomes.

VII. Conclusions

In summary, standards by which evidence in support of implementation of a screening program is judged, should under most circumstances be higher than those supporting the use of an assay for diagnostic purposes. Although in principle the evidence standards applied to genetic or genomic tests should not differ from those used with other types of screening tests, information derived from genome-wide association studies presents unique challenges. Lastly, strong association of a marker with a disease or risk factor alone is insufficient to establish its discriminatory power. Conversely, clinical sensitivity, specificity and positive and negative predictive value help establish the potential effectiveness of a screening marker.