This is a response to the Draft Guidance on "Mobile Medical Applications" (docket # FDA-2011-D-0530; http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm263280.htm). This response is written from the perspective of a physician who has worked on diagnostic decision support software over a 25-year period, resulting in the SimulConsult diagnostic tool.

I write to share my perspectives on these regulatory issues, to which we have paid attention from the start. Although this Guidance is structured in a way that does not affect us directly, the issues raised are important and have wider implications (the Guidance doesn't affect us directly since our software is implemented as a Java applet, a capability not supported by existing mobile platforms).

In the late 1980s the Boston Computer Society convened a joint meeting of its Artificial Intelligence and Legal groups to discuss regulatory issues affecting medical decision support software. The consensus was that software could be expected to be regulated if it worked as an "autopilot", but not if it merely offered advice.

There has long been data in the literature to support such an autopilot versus advice framework. Teach and Shortliffe (1981) studied how doctors interact with decision support software. The results of their survey support the view that doctors do not consider decision support to be an autopilot, they see it as an advice system:

> An ability of a system to explain its advice was thought to be its most important attribute. Second in importance was the ability of a system to understand and update its own knowledge base. Improvement in the cost effectiveness of tests and therapies was also important. Physicians did not think that a system has to display either perfect diagnostic accuracy or perfect treatment planning to be acceptable. (Teach and Shortliffe (1981) "An Analysis of Physician Attitudes Regarding Computer-Based Clinical Consultation Systems". Computers and Biomedical Research 14:542-558).

Public attention has focused on the need to regulate software that can act as an autopilot and lead to significant morbidity and mortality, most notably this year in news coverage of errors in automated radiation therapy (http://www.nytimes.com/2010/01/27/opinion/27wed3.html). The draft regulations follow this spirit in focusing on situations in which there is significant potential for morbidity and mortality. The hard part of the decisions about regulation will be choosing clear criteria that identify situations in which regulation is worthwhile and avoid situations in which regulation has a poor cost-benefit ratio.

A good process for generating such criteria is to examine many individual examples of medical software, decide what regulatory approach seems reasonable for each, and construct a set of clear criteria to systematize such judgments by providing a scoring system for evaluating the need for regulation. Based on my experience with diagnostic decision support software, the following scoring criteria would be reasonable ones for the FDA to use when evaluating a need for regulation of medical software in situations that involve potential for significant morbidity and mortality:

1. **Autopilot**: Consider FDA regulation if the software is used as an autopilot. Note, however, that some software can be like an autopilot for one group of users such as consumers, but not for another group such as medical professionals.
2. **Workable**: Consider FDA regulation for software if quality or testing procedures can be established and implemented in a cost-effective way without markedly curbing innovation and maintenance.
3. **Alternatives**: Consider FDA regulation for software if no good alternatives to regulation exist.

A good way to test such scoring criteria is to examine their effects in individual situations. Some situations seem straightforward. For example, automated radiation therapy software appears to meet all three criteria for regulation. The hard part is to discuss cases that are less clear. I will focus here on the situation with which I am most familiar, diagnostic decision support software.

Criterion 1: Autopilot: As documented by Teach and Shortliffe, diagnostic decision support software is not considered by clinicians as an autopilot. As an example, guided by such research, our SimulConsult tool has several types of screens to explain its advice. Chief among these screens is one we call "Assess disease", which displays the pertinent positive and pertinent negative findings (signs, symptoms or lab tests) in the patient and compares the patient information to what is known about the disease.

The Assess Disease screen shown above illustrates a good match for the patient's findings with the disease MPS I. The patient's pertinent positive findings have reasonable frequency in the disease, as shown by lots of black in the top bars, illustrating that these findings have onset in the disease during the indicated time periods. Furthermore, the pertinent negatives have little frequency, as shown by very little black in the bottom bars. Clinicians use these graphical

displays to help evaluate the various diseases in the differential diagnosis, using the "Scroll in order of differential" buttons at the top right of the screen to scroll through relevant diseases. Clinicians evaluate such information by using their experience, by using narrative material about the disease, and by using lab testing. Access to such validating resources from the Tip links in the yellow area in the figure above is a key part of the decision support. Such access includes links to sites about lab testing, links to narrative articles, as well as a button to the Online Mendelian Inheritance in Man (OMIM) information. SimulConsult is available only by medical professionals because of uncertainties as to whether non-clinicians would have the background and access to information and laboratory testing to use the information in the software as advice, rather than a final answer, and because of uncertainties as to whether consumers would make reliable judgments about findings to input.

Another illustration of clinicians not considering such software to be an autopilot is the design of an ongoing study of SimulConsult's efficacy funded by the National Library of Medicine (NLM). The study compares how clinicians do on diagnostic cases before and after using the software. The study does not focus on how the software itself "performs" because the software is intended as an advice system, not an autopilot. The study is still ongoing, but, for the purposes of this discussion we will assume that such software is capable of significant reductions in diagnostic error and that clinicians don't parrot back decision support advice on diagnosis or workup; instead they take such advice as suggestions to consider.

Criterion 2: Workable: The NLM study, at a cost of ~$378,000, is running tests on 40 cases out of the >2,650 diagnoses in SimulConsult's curated database of neurological and genetic diseases. Such a study is nowhere near as complete as studies of the effects of a drug. Not only is the study examining less than 2% of the diagnoses in the database, but since each patient with a particular diagnosis may have different collections of findings, the scenarios being tested are a very tiny sampling of scenarios likely to be encountered by the software. Furthermore, the NLM study explicitly includes a refinement step in which expert clinicians use the software and its many explanatory screens to look at individual cases and make evidence-based changes to the tool's open database, using screens such as the Define Point screen below:

Define Point — Joint or ligamentous laxity / hyperextensibility / hyperflexibil[...]   Help

Kabuki syndrome   Reset to saved

Lifetime frequency: ○ Sample ● Range ○ Number ○ Verbal ○ Estimate   Reset to normal

50.0 % to 75.0 % (recorded as 62.5%)

Adam M et al. (2011) GeneReviews Kabuki Syndrome   El-Hattab   Segal

Time distribution: ○ Sample ○ Number ● Estimate   Tie to disease onset

Onset: Zero Sharpen Blur Scale to 100%   Total = 100%

10 %	20 %	40 %	20 %	10 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
1w	1m	3m	6m	1y	3y	6y	10	15	25	40	60	80
<2w	2-6w	6-18w	4.5-9m	9m-2y	2-4	4-8	8-12	13-20	20-30	30-50	50-70	>70

Offset: Zero Sharpen Blur Scale to 100%   Total = 60%

0 %	0 %	0 %	0 %	0 %	5 %	10 %	30 %	10 %	5 %	0 %	0 %	0 %
1w	1m	3m	6m	1y	3y	6y	10	15	25	40	60	80
<2w	2-6w	6-18w	4.5-9m	9m-2y	2-4	4-8	8-12	13-20	20-30	30-50	50-70	>70

Adam M et al. (2011) GeneReviews Kabuki Syndrome   El-Hattab   Segal

Cancel   Finding Profile   Disease Profile   Point using another finding   Finish

This case-based refinement step is part of a three-step editorial refinement process, involving case-based, disease-based and finding-based approach. The result is that the database changes many times per month, and currently has >56,000 data points such as the one illustrated above. As a result of these detailed processes to collect evidence-based information, the database is a moving target that changes many times per month. Even if one were to do frequent expensive studies of the effects of the advice of such a system, the results would be informative about only a tiny fraction of the database, and the database itself changes much more frequently than studies could be done, and the studies themselves create new changes. Thus, rigorous FDA Class III regulation that might be applied to a drug would not be workable for such a tool. This is not an argument against doing studies of software and associated databases; indeed SimulConsult urged the NLM to fund such a study of its software. Instead, it is a caution that even those of us who value detailed studies are concerned about mandating such studies because of the costs involved and the difficulty studying a tool that is a moving target.

Other approaches to regulation are possible, such as FDA Class I regulation to institute registration of decision support software and filing of reports detailing procedures for quality control. Although that may be fine in many instances, we need to think more widely about medical information resources and how they may be affected by such rules. One example is IBM's "Watson" software, validated initially in the context of the game show "Jeopardy", but with medical applications being explored by IBM (http://www.technologyreview.com/computing/32427/). Little has been said publicly about IBM's plans, but in writing regulations we need to keep in mind that even the largest of companies can be spooked by uncertainties created by regulation. For tools such as SimulConsult that take an explicitly evidence-based approach, and solicit input from a large

community of clinicians using screens such as the Define Point screen illustrated above, and use peer review of information before publication, one could design an FDA Class I quality improvement procedure to correct for missing data. However, it could be far more challenging to do so for a tool such as Watson, which scans vast amounts of narrative material using natural language processing and uses general rules to process the information. When considering regulations it is important to consider the effects not only on existing tools but on potential tools that could be affected by regulation.

Regulatory issues are an even more important factor for small companies, which may avoid the decision support area entirely or be shunned by investors because of regulatory concerns. Such an outcome would be an especially unfortunate outcome as we approach Stages II & III of "Meaningful Use" of electronic health records, in which clinical decision support are expected to be a core driver of higher quality and lower costs (Health Information Technology Policy Committee, 2011, Meaningful Use Workgroup Request for Comments Regarding Meaningful Use Stage 2. http://healthit.hhs.gov/media/faca/MU_RFC%20_2011-01-12_final.pdf). Such concerns about unintended effects of regulation also affect small businesses applying for SBIR grants, since reviewers of grant proposals may be spooked by the regulatory risks and avoid funding clinical decision support software. Even academic researchers have such concerns about the effects of regulation on funding of grant proposals. Indeed, there has been much discussion of such concerns among businesses of all sizes, medical informatics departments and investors. Even an intention of the FDA to require a mild type of FDA Class I listing could be understood loosely to constitute an amorphous regulatory risk. Such concerns, even if they have little relation to the FDA's actual intentions, may adversely affect innovation. Companies such as SimulConsult might not have been founded if the regulatory climate years ago had offered too many hurdles.

Criterion 3: Alternatives: An FDA Class I listing requirement would offer some useful information. Such information may be particularly useful for products released to consumers, but it is not likely that such information would add much information useful to clinicians. As an example, a non-governmental organization, the Health on the Net Foundation (HON; http://www.hon.ch/) provides a service of certifying quality of thousands of healthcare web sites. Certification is particularly important for consumer sites since consumers are not in a good position to evaluate health information. However, SimulConsult, with a diagnostic tool available only to clinicians, elected not to go through the costs of the HON certification procedure since we believe that far better measures of quality are available to clinicians, and constitute the means by which clinicians make decision to use decision support tools:

- "Word of mouth" referrals are crucial to adoption of decision support tools. For SimulConsult, such referrals have included discussions on physician listservs, peer-reviewed journal articles, product reviews in journals, and invited talks by SimulConsult clinicians at conferences and medical centers.
- One of the top NIH-funded information resources, GeneReviews, began in December 2010 to link to SimulConsult from the differential diagnosis sections of its articles, for example including the following language on http://www.ncbi.nlm.nih.gov/books/NBK1162/ "For a patient-specific 'simultaneous consult' related to this disorder, go to SimulConsult, an interactive diagnostic

decision support software tool that provides differential diagnoses based on patient findings". The decision by GeneReviews to adopt this approach was based on extensive discussions with many clinicians, many of whom had been asking for such functionality to be made part of GeneReviews. (Links from SimulConsult to GeneReviews and back are a form of interoperability done without any financial arrangements, part of a wider theme of interoperability discussed by Segal and Leber (2011) "The Impact of Computer Resources on Child Neurology" in "Pediatric Neurology: Principles and Practice", Swaiman KF et al., editors, 5th edition, Mosby).

- The Child Neurology Society uses SimulConsult in the case-based education it provides for the training of residents (www.childneurologysociety.org/education/casestudies/).
- A leading textbook asks on its first page "With the explosion of electronic information such as found on PubMed and the availability of freely accessible diagnostic software such as SimulConsult one might ask what possible use there is for a handbook like this" (King and Stephenson 2009, A Handbook of Neurological Investigations in Children. 2009, Wiley-Blackwell) and then goes on to highlight its focus on the importance of skills of examination and test interpretation.

To clinicians, such "web of trust" of community endorsements offers much more information than either an FDA Class I listing or an HON certification. Clinicians will be even more impressed with controlled studies such as the NLM study. Of course, evaluations that went beyond the tiny sample of possible clinical scenarios analyzed in the NLM study would be even better, but as discussed under the Workability criterion, such studies would be very expensive and could serve as a barrier to market entry if required for new entrants, even if the studies are funded.

In summary, the 3 scoring criteria enunciate here suggest the following about use of SimulConsult by clinicians:

1. **Autopilot**: SimulConsult is not designed to be an autopilot, and all available evidence suggests it is not used as such by clinicians.
2. **Workable**: FDA Class III regulation would not be workable. FDA Class I regulation could be workable, but superior alternatives exist. Furthermore, even mandating FDA Class I regulation, as benign as it seems, carries a very real danger of spooking the producers of such software, investors, granting agencies and other partners.
3. **Alternatives**: The "web of trust" of community endorsements of decision support software already provides information that clinicians value more than information from FDA Class I reporting or HON certification.

Clearly these 3 scoring criteria would produce different answers for different types of software, and it is likely that others more familiar with other software will be able to improve on the effort here to enunciate 3 scoring criteria for regulation. But the following approaches would be important:

- When extending regulation, it makes sense to begin with cases for which all the criteria point towards regulation.

- For cases in which all criteria point away from regulation, it would reduce the degree to which the marketplace is spooked by the prospect of regulation to make clear that these are cases for which the FDA is *explicitly* not planning regulation.

The importance of getting this right goes far beyond the regulatory issues in themselves. The core financial issue facing the country is the rise in healthcare costs. Many experts believe that decision support software will be a key tool for empowering doctors to change their practice to reduce costs and improve quality. This is a crucial area not to spook with regulation in the instances in which the case for regulation is not compelling.

Michael Segal MD PhD
Founder and Chief Scientist, SimulConsult