C.F. Aliferis

Section of Medical Informatics,
Department of Medicine,
Pittsburgh, U.S.A.

Synopsis

Decision Support Systems

At the core of medical informatics lies a fundamental premise: The enterprise of medicine *can and should* evolve to a highly rational and systematic process with well-defined components, that operate on the basis of sophisticated methods aiming at the optimization of precisely defined outcome goals within the constraints of pragmatic resource limitations.

Perhaps the reason why Medical Decision Support Systems (MDSSs) have attracted so much research attention throughout the history of medical informatics, is that they serve as testbeds for empirically investigating the feasibility of achieving this "Holy Grail" of a rigorous rational/analytic approach to medical discovery, training, and practice. In fact it could be argued that there cannot be a better test for the success of any attempt for structuring and clarifying the medical endeavor than our ability to devise (computational) artifacts that enable, or symbiotically enhance, optimization in both performance (i.e., health outcomes) and cost-effectiveness dimensions.

As a result of this intense research attention, the MDSS field has undergone significant evolutionary steps since the early days of medical informatics. Researchers now have substantial theoretical and engineering tools for, and comprehend fully the importance of:

a. designing systems using *proper architectures* (which enable/facili-

tate efficiency, soundness, expandability, and component reusability and shareability),

b. obtaining and maintaining *high-quality knowledge and data* sources to instantiate their theoretical models,

c. *evaluating* their systems using rigorous methodologies,

d. *integrating the MDSSs into the health-care setting*, and

e. distinguishing between the "*Greek Oracle*" and "*Symbiotic*" models of decision-support and employing MDSS to support the clinical decision tasks rather than aiming at *replacing the health-care provider* [1,2].

As such, MDSS development is a much more mature endeavor, with stronger emphasis given to the engineering rather than the theoretical side, compared to 30 or even 15 years ago.

Partially as a consequence of engineering maturity, increasing numbers of MDSS papers (including several in this section) are nowadays *published in clinical journals*. Presenting quality MDSS work to broader clinical audiences is an objective measure of the field's success in terms of attracting the interest of the broader medical community, and in terms of (eventually) meeting clinical standards for integration of MDSSs in everyday practice.

Another important trend is that of *employing Machine Learning (ML)* representations and algorithms to over-

come weak domain theories, and cumbersome, time-prohibitive, expensive and unreliable human-expert knowledge acquisition. In fact, all major formal representations that have been used for MDSS development (rules, belief networks, neural networks, decision trees) now have widely-available and well-understood learning algorithms, most of them in widespread use [3,4]. ML methods can also help us understand domain properties and augment the MDSS development process even when the latter is carried out by drawing on human expertise.

Finally despite the abundance of practical and stable tools, *theoretical investigations* are still an important part of cutting-edge research. Two papers in this section constitute theoretical efforts that aim at deepening our understanding of sound MDSS design and evaluation. To complete the "bird's eye view" of the contributions in this section, we note that evaluation is emphasized greatly in contemporary MDSSs development, and most papers here have a strong evaluation focus.

To highlight and examine better these trends in the following discussion the *applied* from the *theoretical* papers are separated, and the former are further classified as being *machine-learning based* or *expert-knowledge based*, and the latter as *pertaining to design* or *evaluation*. In the rest of this synopsis the author attempts to present the most interesting contribu-

tions of each paper, as well as broader difficulties/challenges (using the papers as starting points).

Applied Investigations

Machine-Learning-based MDSSs

In their paper, Ortiz and colleagues [5] present a neural network (NN) model for one-year mortality prognosis of patients with heart failure. They compare the model performance against discriminant function analysis, clinical heuristic rules and a rule-learning program. They experiment with several different NN candidates and evaluate their performance against the other methods. They find that the best NNs are reliable and accurate predictors for predicting mortality.

Briggati and colleagues [6] use NNs to identify (early) glaucoma. They compare several NN architectures besides the standard feed-forward back-propagation model, and separate the best ones. They also include comparisons to experts and discriminant function analysis. Their results are very encouraging for the NN models. The authors also examine the importance of structural versus functional variables as predictors of glaucoma and conclude that the latter are more predictive. This is an important function of ML analyses, namely the ability to distinguish between the variables that are essential for a classification task, and the ones that are not. The potential for time and cost savings of such results may prove significant when they are transferred to clinical practice.

Pesonen and colleagues [7] carry out a similar comparison of alternative NN architectures/learning algorithms, this time in the domain of acute appendicitis diagnosis. The back-propagation and learning quantization algorithms prevail in their experiments, as they did in Briggati et al. The authors also try to identify the set of parameters that is

most informative and conclude that clinical signs are superior to clinical history. Their model has a substantial number of cases to train with, and the ones with missing values are excluded. It is possible that this may have led to some selection bias. The authors also experiment with two different versions of the input variables, the first being the original values and the second clinically-abstracted variables. No significant difference is discovered.

In the final ML-based paper of this section, Nussbaum and colleagues address the problem of simulating lumbar muscle recruitment using NNs methods [8]. Although the topic of this paper is not directly focused on MDSS development, the methods employed, as well as the results obtained, are clearly pertinent to MDSSs. Moreover the paper is quite innovative both in terms of a novel application domain and in terms of comparisons made against competing *optimization-based methods*. The authors also found that the choice of network parameters is not critical except for extreme values of the parameters. Based on their findings, the authors conclude that their model "supports the hypothesis that muscle recruitment is controlled by a central pattern generator".

As a general comment, tradition dictates that the currently most widely accepted methods for medical and epidemiological quantitative analysis are biostatistical ones. Therefore any effort to build machine-learning-based MDSSs will likely be more convincing (at least for the present and the near future) to clinicians and statisticians/analysts if it includes comparisons to standard statistical techniques. The authors of the presented papers clearly recognize this necessity. Unfortunately these comparisons (as well as comparisons to alternative ML methods) are associated to a series of methodological and technical difficulties however. For example, the statistics for

comparison of competing classifiers or classifying algorithms, even in a single domain and data set, are fairly undeveloped [9]. In addition, if one wants to apply multiple alternative statistical and ML methods to the same problem, cross-validation (or equivalent techniques) are important to negate the effects of bad random splits of data into train and test sets, and overfitting. In many cases, more than one nested cross-validation layers will be needed for tuning the ML algorithms parameters, convoluting thus the experimental designs and the associated inferential statistics.

Another concern is that for the comparison to be unbiased, each ML method should be employed by researchers who are well familiar with the individual techniques (many ML techniques cannot be operated fully automatically). This generates problems of *accessibility of expertise as well as potential bias* because of different skill levels of the individual workers (within their particular ML specialty). Many more problems exist that are outside the scope of this synopsis, however it is important to realize that *most of these problems will appear in the context of any significant effort to develop realistic-scope ML-based MDSSs*. For example in the paper of Ortiz et al., a great deal of effort is given to developing the NN model, while the rule-based classifiers are built using a 1-rule classification model. This is not the case with modern rule-based learners (and certainly not with human knowledge acquisition), which typically produce, evaluate, and process rule-sets with hundreds of rules [10]. Also in 2 papers discriminant function analysis is used while logic models might have been the first choice of modern statistical practice (as a note, logistic regression with discrete independent variables and enough interaction terms can capture non-linear discrete conditional probability distributions -- contrary to the

concerns about non-linearity expressed in one paper). Similarly the issues of statistical comparisons among the techniques, as well as repeated and nested cross validation are also relevant to all papers. These comments are *not* a criticism of the sound work presented in this subsection, as they only serve to indicate the complexities and challenges of properly developing ML-based MDSSs. For a comparison of ML methods that makes explicit and tries to address some of these broad issues see also [11].

Expert Knowledge and Literature-based MDSSs

Balas and colleagues [12] present a MDSS that aims at helping physicians and managers/organizations achieve conformance to sound clinical guidelines as well as analyzing the physicians' performance. The approach employed ("*direct mail*" model) shares, as implemented, a number of similarities to other efforts which are aimed at bringing solicited or unsolicited information to the point of healthcare. The latter research includes reminders [13], automatic document retrieval [14], and MDSSs for patient-specific adaptation of literature [15, 16]. It is differentiated from these efforts however in that (a) is targeted to individual health-care providers, rather than patients or provider *acts*, (b) addresses goals at the health-care organization level, and (c) is heavily based on peer comparison. The methodology seems a promising tool for sound evidence-based medical practice. On the other hand, significant precautions seem necessary (especially in view of the growing concerns about managed health-care practices) to *tailor the guidelines* to individual populations of patients that may be inherently different than the ones at which the guidelines are aiming. This and the issue of developing enough practice guidelines for precisely recognizable clinical settings/contexts seem to be the biggest chal-

lenges for the success of such systems.

Post and colleagues [17] present the design and second-phase evaluation of FreeCall, a MDSS aiming at increasing the cost-effectiveness of ambulance dispatching decisions in the context of the Dutch Health-care System. FreeCall is a Knowledge- Based system that implements a *hybrid physiological and quasi-probabilistic knowledge* model to reach decisions about either admitting patients who call about thoraco-abdominal and consciousness problems, or referring them to the GP. The MDSS was successful in reducing the number of dispatching ambulances for non-urgent conditions, and demonstrated a slightly higher threshold for admitting truly urgent cases relative to human dispatchers by missing 1 urgent case out of 14. A point to note is that it is not clear in the analysis presented whether these differences are non-random. Moreover, as the authors point out, explicit cost-effectiveness criteria need to be devised for this decision problem and used to further validate (as well as enhance) the system's performance.

Keith and colleagues [18], describe a large multi-center study of an MDSS for managing labor using information from the cardiotocogram. They compare their system against 17 experts and reach interesting conclusions about the system and also about the experts' attitudes towards the information presented to them, and their decision-making profiles regarding cesarean sections. The system's performance is such that it allows the authors to suggest that *there is significant information in the cardiotocogram that is not being used appropriately by physicians*. This is an important conclusion in its own right. An interesting methodological point is that this evaluation study has a *strong temporal component*, since decisions can be made at any point during labor. The researchers simulated the flow of in-

formation to the subjects (experts) as well as the system and examined (among other performance metrics) *measures of timing agreement*. The system exhibited excellent performance both in terms of avoiding unnecessary intervention and in terms of identifying cases requiring intervention.

Theoretical Foundations

MDSS Design Theory and Methods

Johansson and Timpka [19] bring the methodological perspective of *requirements engineering* to the study of MDSS design and development methods. Using three converging observational studies with MDSS-expert subjects they develop *quality requirements* for MDSS development/design methods and examine the relationships among them as well as among those and subjects needs. Because the paper is written from the perspective of a methodology largely unknown to many in the MDSS audience, and due to lack of several operational details as well as to the usage of fairly abstract language, the paper may be somewhat difficult to follow at a first reading. However, both from a methodological and a results viewpoint it deserves attention since:

- a. It provides an example of a carefully executed analysis for determining systems methods requirements at a level of abstraction much broader than what is usually encountered in the literature, and gives several references to similar and supporting methods work.
- b. Its results are interesting. Among other findings for example, Johansson and Timpka found that a MDSS method should "be custom tailored to health-care", "help to focus on the clinical decision-making situation being the foundation of the

development work", and "help maintain the goal for the development and product as well as the goals in the clinical environment".

Insights like these may prove useful for developing effective design guidelines for large-scale MDSS design and engineering efforts which would prevent several ex post facto problem identifications. Adopting early on a feel for the needs of both the providers and customers of MDSSs using a principled (if not strictly experimental) approach, may prove to be a valuable exercise for any MDSS development methodology that exceeds anything but the smallest systems. Clearly much research, as the authors themselves advise, needs to be done in investigating these and several more questions discussed in the paper.

MDSS Evaluation Theory and Methods

Kahn and colleagues [20] introduce *statistical process control methods in MDSS performance monitoring*. These methods were originally developed for quality monitoring of industrial processes. The paper addresses specifically the problem of obtaining multiple, sequential post-deployment evaluations of MDSSs in dynamic changing domains/health-care environments. That is, evaluation after what traditionally is considered the "final" formal MDSS evaluation stage (i.e., RCT study or equivalent at the target site). Such ongoing evaluation is extremely important when the parameters of the domain and/or the system usage are evolving in time, and when controlled experiments are too expensive to be repeated frequently (as is almost always the case). The authors introduce and explain clearly several concepts, many by means of an example evaluation of their GermWatcher system. Moreover they discuss several aspects of how to select appropriate monitoring criteria and methodologies. The paper's motivation is clear

and compelling, and the methods presented are simple and tested (albeit in different fields). Whether statistical process control will become a standard methodological tool for evaluation after MDSS deployment, remains to be seen. However it seems safe to conclude at this (early) stage that its concepts deserve to be investigated as a sound starting point towards solving the problem of ongoing MDSS evaluation.

Conclusion

The papers in the 1997 Yearbook of Medical Informatics section on MDSSs, give support to the notion that the field is progressing with intellectual vigor and engineering maturity. Important problems (several of which were highlighted by this author during the discussion of the papers) notwithstanding, research contributions like the ones reported here are important steps towards:

- a. *Expanding the boundaries and clarity of our theoretical understanding of the field* --both in terms of design and evaluation.
- b. *Meeting the practical needs for high-quality systems*. Such systems are becoming increasingly more acceptable and useful at the clinical end because of their tackling important tasks, and by virtue of their sound design, convincing evaluations, and appropriate integration into the health-care setting.

Acknowledgment: The author thanks Dr. W. Hogan for reading this synopsis and offering useful comments.

References

1. Miller RA. Medical diagnostic decision support systems - past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc* 1994;1:8-27.
2. Shortliffe EH. The adolescence of AI in

- medicine: Will the field come of age in the '90s? *Artif Intell Med* 1993;5:93-106.
3. Weiss SM, Kulikowski CA. *Computer Systems that Learn*. San Francisco, CA: Morgan Kaufmann Publishers, 1991.
4. Mitchell, TM. *Machine Learning*. New York: McGraw Hill, 1997.
5. Ortiz J, Cheffer CGM, Silva CES, Sabbatini RME. One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data. *J Am Coll Cardiol* 1995;26:1586-93.
6. Brigatti L, Hoffman D, Caprioli J. Neural networks to identify glaucoma with structural and functional measurements. *Am J Ophthalmol* 1996;121:511-21.
7. Pesonen E, Eskelinen M, Juhola M. Comparison of different neural network algorithms in the diagnosis of acute appendicitis. *Int J of Biomed Comput* 1996;40:227-33.
8. Nussbaum MA, Chaffin DB, Martin BJ. A back-propagation neural network model of lumbar muscle recruitment during moderate static exertions. *J Biomech* 1995;28:1015-24.
9. Dietterich TG. Statistical tests for comparing supervised classification learning algorithms. (Submitted. Preprint available at: <http://www.CS.ORST.EDU/~tgd/cv/jr.html>).
10. Ambrosino R. *The development and evaluation of a method for rule post-processing*. Report SMI-95-02, Section of Medical Informatics, Department of Medicine, University of Pittsburgh, 1995.
11. Cooper GF, Aliferis CF, Ambrosino R, Aronis J, Buchanan BG, Caruana R, Fine, MJ, Glymour C, Gordon G, Hanusa BH, Janosky JE, Meek C, Mitchell T, Richardson T, Spirtes P. An evaluation of machine learning methods for predicting pneumonia mortality. *Artif Intell Med* 1997;9:107-38.
12. Balas EA, Zong Rong Li, Spencer DC, Jaffrey F, Brent E, Mitchell JA. An expert system for performance-based direct delivery of published clinical evidence. *J Am Med Inform Assoc* 1996;3:56-65.
13. McDonald CJ, Hui SL, Smith DM, et al. Reminders to physicians from an introspective computer medical record. *Ann Intern Med* 1984;100:130-8.
14. Wagner MM, Cooper GF. Evaluation of a belief-network-based reminder system that learns from utility feedback. *J Am Med Inform Assoc* 1996;3(S):666-72.
15. Rennels G. *A computational model of reasoning from the clinical literature*. New York: Springer-Verlag, 1987.
16. Lehmann H. *A Bayesian computer-based approach to the physicians' use of the clinical research literature*. Stanford University: PhD dissertation, 1992.

17. Post WM, Koster W, Sramek M, Schreiner G, Zocca V, de Vries B. FreeCall, a system for emergency-call-handling support. *Meth Inform Med* 1996;35:242-55.
18. Keith RDF, Beckley S, Garibaldi JM, Westgate JA, Ifeachor EC, Greene KR. A multicentre comparative study of 17 experts and an intelligent computer system for managing labour using the cardiotocogram. *Br J Obstet Gynaec* 1995; 102:688-700.
19. Johansson M, Timpka T. Quality functions for requirements engineering in system development methods. *Med Inform* 1996;21:133-45.
20. Kahn MG, Bailey TC, Steib SA, Fraser VJ, Dunagan WC. Statistical process control methods for expert system performance monitoring. *J Am Med Inform Assoc* 1996;3:258-69.

Address of the author:

C.F. Aliferis
Section of Medical Informatics,
Department of Medicine,
Suite 8084 Forbes Tower, 200 Lothrop St.,
Pittsburgh, PA 15213-2582, USA.
e-mail: cons@cbmi.upmc.edu