Received: 4 March 2018 | Revised: 2 July 2018 | Accepted: 3 July 2018

DOI: 10.1002/lrh2.10065

COMMENTARY

Learning Health Systems

Desiderata for sharable computable biomedical knowledge for learning health systems

Harold P. Lehmann[1] | **Stephen M. Downs**[2]

[1] Johns Hopkins University, Baltimore, Maryland

[2] Indiana University, Indianapolis, Indiana

Correspondence
Harold Lehmann, Johns Hopkins University, E Monument St, 1-201, Baltimore, MD 21205.
Email: lehmann@jhmi.edu

Funding information
HPL: PCORI, Grant/Award Number: CDRN-1306-04912

Abstract

In this commentary, we work out the specific desired functions required for sharing knowledge objects (based on statistical models) presumably to be used for clinical decision support derived from a learning health system, and, in so doing, discuss the implications for novel knowledge architectures. We will demonstrate how decision models, implemented as influence diagrams, satisfy the desiderata. The desiderata include locally validate discrimination, locally validate calibration, locally recalculate thresholds by incorporating local preferences, provide explanation, enable monitoring, enable debiasing, account for generalizability, account for semantic uncertainty, shall be findable, and others as necessary and proper. We demonstrate how formal decision models, especially when implemented as influence diagrams based on Bayesian networks, support both the knowledge artifact itself (the "primary decision") and the "meta-decision" of whether to deploy the knowledge artifact. We close with a research and development agenda to put this framework into place.

KEYWORDS

Bayesian analysis, decision analysis, decision support, knowledge engineering, predictive modeling

1 | THE PROBLEM

"Learning health systems" (LHS) create knowledge, but they also create knowledge objects (KOs)[1]: rules, guidelines, predictive models, and the like.[2] Applying what we learn in 1 complex adaptive system[3] to another is fraught with dangers: How can 1 complex health care environment resemble another? And yet, share we must. Researchers of clinical prediction rules recommend a well-articulated sequence of evidence collection leading to broad dissemination.[4] However, we contend that the enthusiasm for modern computation[2] and the demand for Big Data[5] to provide quicker results[6] lead to the expectation that LHS KOs shall be shared quickly and broadly.[7]

Making KOs accessible and reusable has motivated much work in decision support, including at the very least Arden Syntax,[8] FHIR,[9] and CDS Hooks.[10] The BD2K community applies FORCE11's acronym, FAIR, to this goal: Findable, Accessible, Interoperable, and Reusable.[11] These features of KOs shared among LHS would seem to be the minimum requirement.

The KOs that are shared among clinical decision support systems (CDSs) for electronic health records (EHRs) generally result from long processes, comprising evidence collection over several years, review by the community of practitioners and experts, and achievement of consensus.[12] As opposed to such complex sociotechnical processes, the knowledge gained from LHS is derived from observational data and is generally much earlier in the epistemological process than traditional CDS rules. They are, thus, closer to the data from which they were derived. Our claim is that those who share LHS KOs have a responsibility to maintain the link between the knowledge and source

data much differently from traditional CDS. The resulting knowledge representation (KR) will need to be different from CDS KR as well. The frameworks used to implement sharing of CDSs are wrongly applied to LHS.

In this commentary, we will work out the specific desired functions expected of LHS knowledge sharing, and, in so doing, discuss the implications for novel knowledge architectures. We will demonstrate how *influence diagrams* satisfy the desiderata. Influence diagrams are a specific format for representing decision analytic or cost-effectiveness models as directed acyclic graphs. An important conclusion of this commentary is that decision modeling is integral to the process of sharing LHS KOs. We use the threshold theory of decision making as an organizing principle, where the optimal decision is suggested by whether the mean belief (estimate) of a quantity of concern, like mortality risk, is above a threshold determined by the decision maker's assessment of the tradeoff between harms and benefits of action or inaction.[13,14]

We acknowledge 2 levels of decisions involved in the use of these KOs: the decision implicit in the artifact and the institutional decision about whether to use the artifact (the "meta-decision"). Our desiderata concern both.

2 | THE DESIDERATA

The following are desiderata for making a data-based model usable at a local institution remote from its site of creation or validation. We recognize that a KO library can be built and not meet all or even any of these items; hence, we cannot call them "requirements." "Local validation," a recurring component of the desiderata, includes local representation, recalculations, and assessment of the newly implemented model's characteristics. We use as our use case the question of blood lead screening in children: What should the local rule be?

2.1 | LHS models shall enable local validation of discrimination

A model must discriminate cases from noncases, whether a "case" is a diagnosis (eg, high blood lead level) or a health state, such as intellectual achievement. The usual measure of discrimination is the receiver operating characteristic (ROC) curve's area under the curve (AUC). Figure 1 shows 1 author's (SMD) experience in describing models based on either observed data or expert opinion.[15] While higher numbers are better, left unstated in almost every report of a model's performance is what threshold AUC is considered acceptable (for the meta-decision). Moreover, identifying the optimal cutoff point (false positive rate [FPR] and true positive rate [TPR]) on the curve for a local implementation requires decision modeling and valuation of the "cost" of a false positive or false negative.

2.2 | LHS models shall enable local recalibration

If, for the same inputs, 1 model calculates a risk of MI of 30% and, another, 40%, and the patient's threshold for concern is 35%, then that patient will feel advised in diametrically opposite directions. The calibration of risk models varies across instances of use, and depends on the local population.[16] The need to focus on calibration in the neighborhood of the threshold led Walsh and colleagues to offer a novel calibration approach based on its performance near the decision threshold.[17]

2.3 | LHS models shall enable incorporation of local preferences into local calculation of thresholds

Both of the first 2 desiderata raise the notion of thresholding, either for the meta-decision (what maximum FPR and minimum TPR would the clinicians find acceptable) or for the decision itself (act on the

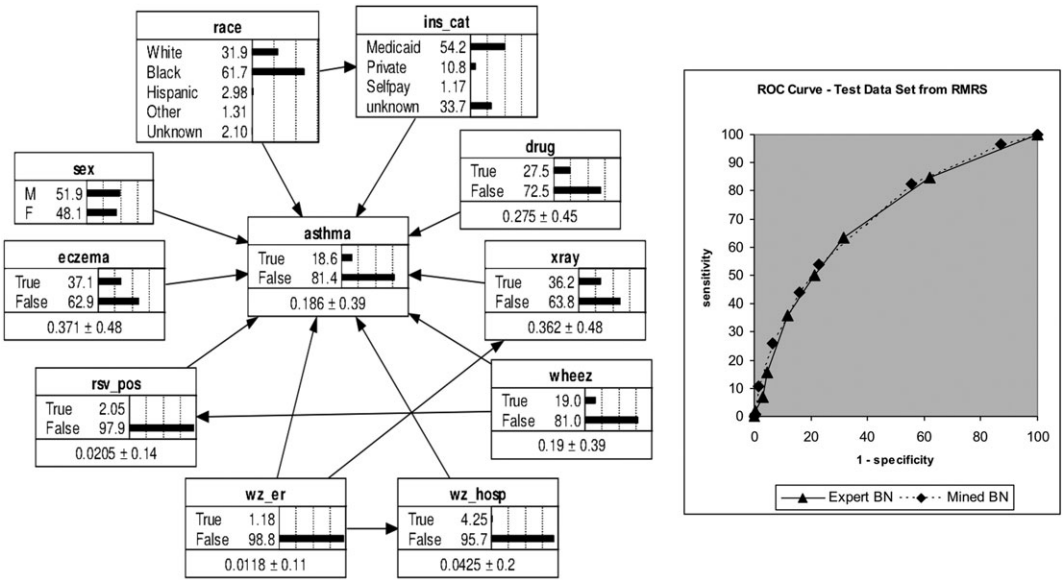

FIGURE 1 Asthma risk. A, Bayesian network (built in Netica™). Rectangles represent probabilistic nodes, the bar graphs, the probability distributions. The arrows between nodes reflect conditional probabilistic relationships (numbers not shown). B, ROC resulting from the Bayesian network. Two models are reflected: one based on probability estimates elicited from experts and the other from data in the RMRS electronic health record system. The ROCs almost overlap each other. The AUC for each is 0.70[15]

mortality risk), yet such thresholds are rarely explicitly articulated, nor methods for their assessment, elicitation, or calculation provided.

In the LHS context, there are at least 3 types of thresholds. For the primary decision, there are 2: risk (probability) thresholds and cutoff levels. The first type of threshold is what risk level is high enough to institute risk-based screening.[18] If the risk is high enough, then action (eg, screening) makes sense; local factors (eg, subgroup compositions) map to a risk level, which lead to the rule: For this locale, you should (not) screen. An example of the second type of risk threshold, the lab test cutoff, is the blood lead level above which action should be taken. These cutoffs are often presented as resulting from a statistical analysis. The history of lead-level cutoffs is a classic example: Back in the 1970s, the CDC recommendation for referral was 25 mcg/dL.[19] In 1981, a decision analysis calculated a threshold for an ROC for a lead test at 35 mcg/dL.[20] The threshold today is 5.[21] The changes in the cutoff reflect research on the long-term neurocognitive changes, including what level of such changes is acceptable. While the cutoff may be expressed as a percentile (today's "97.5%th percentile," justifying the 5 mcg/dL cutoff[21]), the reasoning for this level is left implicit (Figure 2).

As an example of determining a risk-based rule, see Figure 3.

Figure 3 shows how the calculation of expected value (utility) for all possible values of risk suggests when different strategies are recommended by the model. Thus, the same model can give different recommendations, based on local values of key parameters, like base risk of levels above the reference level.

The third type of threshold is the cost-effectiveness threshold, required for the meta-decision. The measure of cost-effectiveness, the incremental cost-effectiveness ratio (ICER), is compared with a maximum willingness to pay, the threshold for this decision problem.[24]

These thresholds are derived from decision modeling. Decision analytic modeling presents a particularly appealing framework for representing knowledge in a LHS KO. Properly designed, a decision model can represent statistical information at an arbitrary level of detail, including representation of the analytic approach and potential biases. Moreover, decision models represent the preferences that are the basis of setting thresholds. The components of a decision model are mnemonically encapsulated by the term You SHOULDT: You is perspective, S, Structure, H, history or who enters the model, O, outcomes of concern, U, uncertainties, L, list of alternatives or options, D, desires or tradeoffs, and T, time horizon. (See Supplement.)

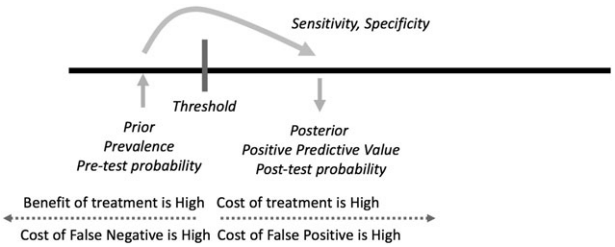

FIGURE 2 Threshold model of decision making. The horizontal line reflects the probability of the outcome driving the decision. The vertical up arrow, what probability the decision maker holds before gaining information; the vertical down arrow, the probability after that information

Decision models establish decision thresholds using expected utility theory, which is predicated on the notion that the preferred option in a decision problem is the one that offers the highest expected (weighted average) utility. Utility is a function on the outcomes of the decision problem that quantifies preferences. Methods for deriving utilities were first devised by von Neumann and Morgenstern in 1947.[25] Working from 5 intuitive axioms of preference under uncertainty, they derived the standard reference gamble that allows the decision maker to assign any outcome the correct preference value (or utility) relative to other outcomes. The resulting utilities have the property described above.

In the intervening decades, numerous methods have been derived that approximate the standard reference gamble and have better psychometric properties and lower cognitive burden.[26] It has become conventional to anchor health utility values on 0 (for death) and 1 (for perfect health).[27] There have also been a number of standard utility scales developed to more easily assign "population" utilities to health states.[28,29]

To illustrate how a decision model might provide the right framework for constructing and using an LHS KO, we consider a decision rule to screen children for lead poisoning. The rules says that if a child has any risk factors for lead poisoning (anemia, siblings with high blood lead levels, or living in a house build before 1960), s/he should be tested with a blood lead level. The rule is illustrated by the algorithm shown in Figure 4.

This decision rule has a great deal more background, which can be represented by the decision model in Figure 5 as an influence diagram based on the last published simulation in the literature, but with added concerns about risk factors for increased blood lead levels.[31] The darker nodes are directly represented in the algorithm. An influence diagram has nodes representing probabilistic, deterministic, decision, and utility concerns.[32,33] In essence, they are Bayesian belief networks[34] with decision and utility nodes added. They thereby encompass both statistical and decision models.[35] The ovals (chance nodes) repres\ent random variables. An arc (or edge) directed toward a random variable indicates another variable on which its probability distribution is conditioned. Thus, blood lead level increases the likelihood of anemia; having a sibling with a high lead level or a home build before 1960 increases the probability of an elevated blood lead, and an elevated blood lead increases the chance of a positive result on a blood lead test.

The rectangular (or decision) node represents the choices from which the decision maker can select, in this case, screening or not screening for lead poisoning. An arc entering a decision node indicates information that will be available at the time the decision is made. The decision maker knows about the risk factors before making a decision to screen.

The diamond node is a value node. It represents the value (utility) for the potential outcomes. Arcs entering this node indicate the variables that affect the utility. The "treat" node is deterministic. In this model, it indicates that if the blood lead result is above 5 mcg/dL, the child will be treated.

Evaluation of the model in Figure 5, as is, will result in the same decision rule as the algorithm in Figure 4.

However, the model makes explicit all of the reasoning and data behind the recommendation. The structure of the model makes clear

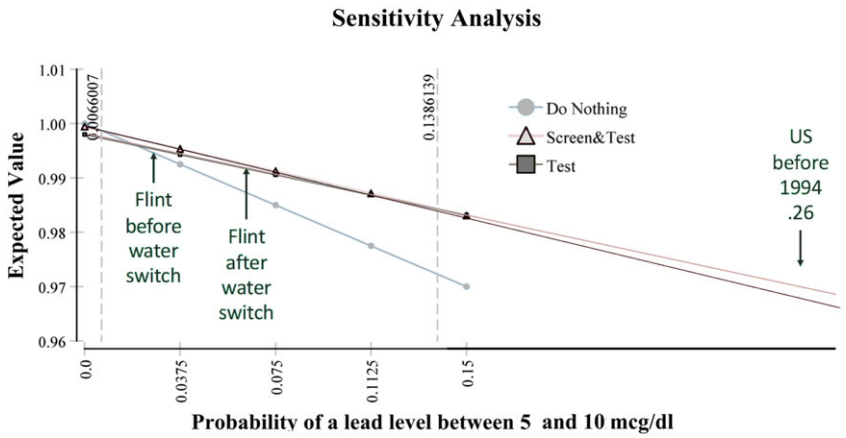

FIGURE 3 Sensitivity analysis for the lead-testing problem. Higher values of expected value are preferred to lower values, so the screen-and-test strategy (△) is preferred between the 2 blood lead risk thresholds of .0066 and .139 mcg/dL. To the left of that interval, the do-nothing strategy (○) is preferred, and to the right, the test strategy (□) is preferred. In 1994, the general US probability was .26,[22] to the right of the interval, so "test" was preferred (and indeed recommended). In the case of Flint, MI, the strategy of screen and test had the highest expected value before and after the water switch.[23] The value of spending resources to get a better estimate of the probability of a lead value between 5 and 10 mcg/dL can is related to the distance between the best and second-best strategies, wherever the locale's current estimate lies on the *x*-axis

how the problem was conceptualized and the probability distributions in each of the chance nodes. Moreover, the statistical methods for deriving the distributions, including confidence limits and corrections for biases, can be represented with additional nodes.[30] The values (utilities) chosen by the KO creators are made explicit

A KO that is presented in this kind of detail can be evaluated and adjusted by users. For example, the prior probability of an elevated blood lead level can vary dramatically from 1 location to another. Value judgments on the utilities of outcomes may vary from place to place.

Elicitation or assessment of preferences required to localize a model can be done qualitatively or quantitatively. Qualitative preferences can be used in the context of sensitivity analysis (see below). Quantitative preferences can be assessed in many different ways, such as standard gambles[36] or multiattribute assessments,[37] both of which are based on formal theories of preference.[38] The analytic hierarchical process[39] and conjoint analysis[40] are examples of methods used that are not tied to an axiomatic basis, but work well enough.

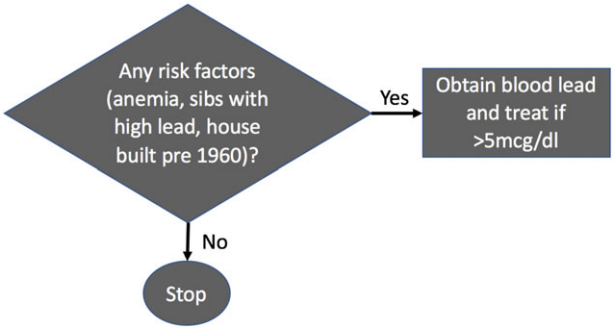

FIGURE 4 An algorithm depicting a decision rule for blood lead screening. If a child has any risk factors for lead poisoning shown, s/he should be tested with a blood lead level. A rule like this is easily implemented in all EHR-based CDSSs

2.4 | LHS models shall provide explanation

"Explanation" covers a wide scope: justification, provenance, and why the model came up with the answer it did in a particular case. Models "justify" themselves in their evidential rigor. Provenance affirms the sources of the data[41]; we need the equivalent for models. Justification and provenance that provide transparency are generally identified in static meta-data.[42] Answering why a model came up in a particular case requires dynamic computation. In the early days of medical expert systems, we learned both that clinicians wanted explanation[43] but rarely sought it out (eg, 5 uses of Infobuttons per user per month[44]), at least in decision support transactions. With complex models, we are finding that clinicians indeed want an explanation of sorts. For instance, Austrian and colleagues provide an interface for sepsis surveillance that indicates with red (vs black) font color which of the patient's finding triggered the alert.[45] Their override rate was only 3.7%, vastly lower than the typical rate of 49% to 96%.[46] Showing what data lead to which results constitutes a modern form of explanation. Decision models in general, and influence diagrams in particular, can generate such explanations.[47]

2.5 | LHS models shall enable local monitoring

Interventions in complex adaptive systems lead to unpredictable results.[48] Only with monitoring will those responsible for the health system know whether the intervention is working and the level and types of harms resulting from the new intervention. Yet monitoring brings its own complications. Take the example of risk-based pediatric lead screening. If the health system decides to implement such screening, then any child not meeting the explicit risk profile (or risk threshold) will not be screened. Monitoring only the blood lead levels available from the clinical lab system means that clinical leaders will no longer be monitoring the entire population of children, just the population of those who screen positive. To assess whether those screening negative are indeed below threshold requires a separate

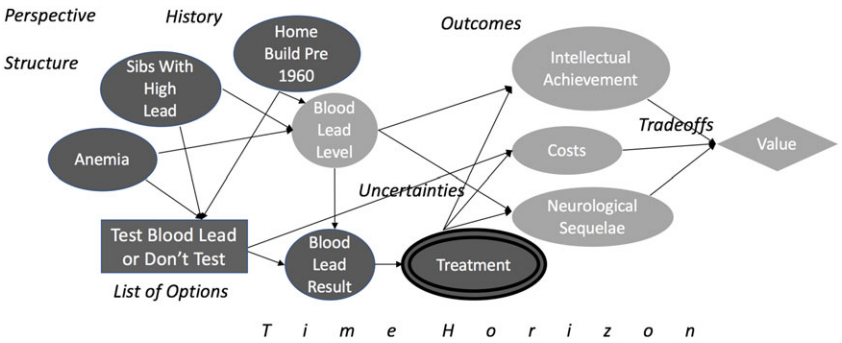

FIGURE 5 Example influence diagram for deciding whether to test a child, or children in general, for lead poisoning. At the point of the decision, whether to test blood lead, the decision maker knows the patient's risk factors (anemia, siblings with high lead level themselves, and whether their home was built before 1960). These risk factors have a probabilistic impact on the actual blood lead level, which is not observed, but whose probabilistic relationship (sensitivity, specificity) to blood lead test result is known. If the measured blood level is above a cutoff (not shown), then treatment occurs certainty (deterministically). The value (utility) to the patient will depend on what the actual blood lead level was, as well as the effectiveness of treatment; this value results from neurological sequelae and costs. The you SHOULDT components are labeled. (See supplement for details)

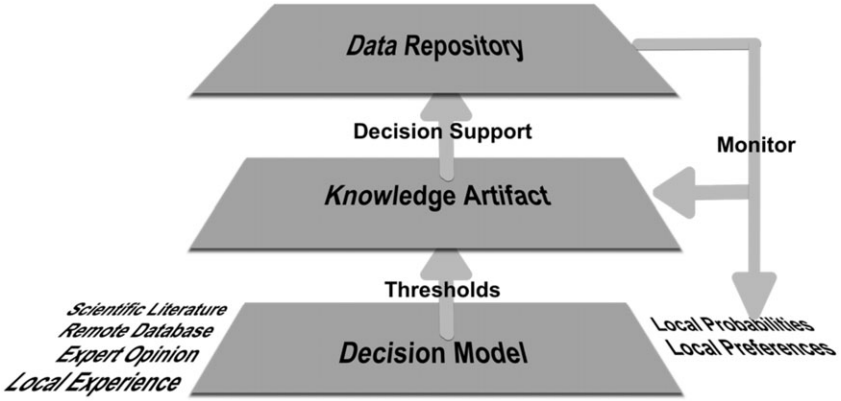

FIGURE 6 The knowledge artifact localization cycle

surveillance strategy. Yet, if the clinical leaders are not aware of this skew brought about by the intervention of risk-based screening, they will have a biased understanding of their population.

Figure 6 summarizes the desiderata to date—the knowledge artifact localization cycle. The figure makes clear that the decision model is where localization should occur; it is at this level that local experience and opinion, as well as local regard for the scientific literature and available remote data, such as registries or public health data, should be incorporated and integrated. With probabilities and utilities specified, the mathematical cranks of utility maximization, sensitivity analysis, and value of information[49] can be turned: Thresholds are determined, and the local version of the knowledge artifact is generated—an artifact assured to be consistent with all the knowledge supplied. That artifact may be turned into decision support—rules, dashboard, guidelines, and other knowledge widgets.[50] In the course of use, data are generated into the data repository (eg, enterprise data warehouse), which data are then used for monitoring. If the thresholds do not change, then the knowledge artifact may be simply updated; if thresholds do change, the knowledge artifact may need to be regenerated. The notion that the decision model is the locus of knowledge maintenance appears in Chapman's and Sonnenberg's monograph.[51]

The decision model can itself be used to determine which parameters need to be monitored: Calculating the value of information for each parameter in the model points to the parameter having the biggest influence on whether the institution is below or above the threshold for action.[52,53]

2.6 | LHS models shall enable debiasing

The notion of biases in monitoring raises the much larger concern for bias in general. Epidemiologists have spent decades articulating different types of bias that analysis of observational (and other) data must take into account. For instance, Sander Greenland has spent gathered his quantitative models in a recent chapter of his in a definitive epidemiology text book.[54] These biases generally fall into 4 types: selection, information, confounder, and execution of intervention.[55] For observational data, like EHR data, authors have identified at least 16 biases. Figure 7 attempts to place where these biases operate, along a spine of successive dependencies of patient populations and cohorts. A 17th bias, semantic uncertainty, arises because of variability in the database query uses.

The monitoring bias identified under Monitoring is the inverse of sick quitter bias: Instead of losing the experience of those who drop out of observation because of the progression of their surveilled illness, it is the "well" patients who are lost to observation. (See Supplement for an estimate of this bias.)

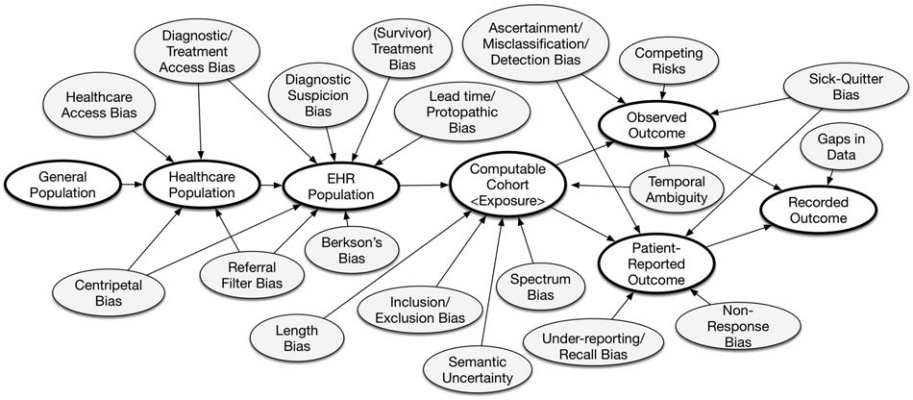

FIGURE 7 Bayesian meta-model for debiasing. Specific biases come from Delgado-Rodriguez[56]

Lehmann and Shachter showed how some of these biases may be represented in Bayesian networks at 2 levels: qualitatively and quantitatively.[57] The qualitative representation is based on the location of bias sources that act at the level at which they appear: population, sample, observed, and reported.

Bayesian networks can also represent the quantitative model. Over the past 20 years, generalized software (OpenBUGs,[58] BRugs,[59] Stan[60]) have been developed that, given a coherent model, along with prior distributions and data, can provide estimates and credible sets (Bayesian confidence intervals) for the parameters in the model, parameters representing both the domain concerns (eg, relationship between blood lead levels and risk factors) and bias concerns (eg, crossovers).

To date, no LHS models have taken such models into account in their analyses or in their dissemination. The recent attention to the informatics of data quality points out the types of errors, if not biases, resulting from the EHR environment.[61] However, these data quality issues are articulated as issues that need to be assessed, not as biases generated by specific observation and documentation mechanisms.[62-64]

2.7 | LHS models shall account for generalizability

To apply an LHS knowledge artifact at a new institution means that the institution must assess whether or how that knowledge applies locally. This concern is not limited to issues of population similarity, but to larger range of sociotechnical factors that speak to the meta-decision: Are the treatments involved available? Can the institution apply them the same way? What barriers would prevent successful implementation?[62-64]

The LHS KO must, "out of the box," tell the prospective new locale how the "knowledge" contained in the KO should be modified, taking local realities into account. The method by which this modification is performed is the essence of generalizability.

In controlled studies, the goal of generalizability is reflected in the appeal to some sort of sampling or other strategy to make the study population as "representative" of the target population as possible. In LHSs, explicit attention to selection biases must be made. There are 2 approaches to addressing this generalizability concern. One is through statistical modeling.[57]

Figure 7's meta-model depicts the levels of concern. The other is more heuristic, calculating the "distance" between the studied population and the target population. Some researchers have worked out heuristics for calculating this distance, and using that distance to change the implications of results, with GIST 2.0 as an example.[65]

2.8 | LHS models shall account for semantic uncertainty

Semantic interoperability has been a goal of much work in informatics.[66] However, analysts are quite aware that applying vocabulary or conceptual standards to their data may amount to stuffing that data into a Procrustean bed,[67] chopping off or stretching out the intention of the data. Ideally, we would have an *epistemic confidence interval*, whose uncertainty would communicate uncertainty, not because of sampling, but because of the semantics-related modeling decisions made along the way, and would communicate the corners cut in so doing. Sensitivity analysis makes sense as an approach to accomplish this communication, but has not been used much, to date.

2.9 | LHS models shall be findable

Findability is the first of the FAIR data principles.[68] One would imagine that this desideratum would precede the rest: After all, before assessing discrimination, calibration, and the rest, one needs the model in hand.

We left this desideratum for last, because we now know the elements we need to "find" the appropriate model: As statistical models, they require ontological concepts for parameters, statistical distributions, relationships among the parameters, the types of outcomes, and the intention of the model at both statistical level and computational levels. As epidemiological models, they require ontological concepts for the design of the study leading to the artifact, for the biases represented, and for the qualitative and quantitative "hooks" for biases not explicitly included. As decision models, they require ontological concepts for attributes of utility. For generalizability, they require ontological concepts important in determining the "distance" from the local dataset to the data of the knowledge artifact.

Some currently available ontologies meet some of these needs: STATO supplies concepts for statistical tests, conditional of

TABLE 1 Desiderata and work needed to be done

Desiderata	Development Work to be Done
1. Discrimination	Measures that take clinical thresholds into account[73,74] Elicitation and articulation of those thresholds Methods for recalculating local discrimination
2. Local recalibration	Application of calibration based on thresholds[17]
3. Thresholds and local preferences	Elicitation, articulation of preferences Local calculation of thresholds
4. Explanation	Deployment
5. Monitoring	Choose variables based on value of information[75]
6. Debiasing	Creation and curation of debiasing models Application of debiasing models
7. Generalizability	Calculation of distance[65] Adding to the knowledge artifact the meta-data required to choose the calculation
8. Semantic uncertainty	Derivation of the epistemic confidence interval
9. Findable	Articulation of the full ontology required to index a knowledge artifact at all its multiple levels Tagging KO with that ontology
10. Other commandments as necessary and proper	Continuous monitoring and improvement of these desiderata

application, probability distributions, and experimental designs.[69] The OCRe ontology addresses study designs,[70] and CDISC, study execution.[71]

2.10 | LHS models shall enable future properties as necessary and proper

No listing of desiderata can be complete in this early phase of KO development and sharing. As such, we reserve an "elastic clause" to enable additions as the community gains experience.[72] Two issues not already addressed, for instance, is that models should explain how their components (existing or new ones) should be integrated together and that their entry in the shared library should accrete the experience of those who have used the models.

3 | DISCUSSION

In this commentary, we have developed 10 desiderata for sharing KOs based on LHS experience, and aimed at reapplication in other locations.

Table 1 lists the desiderata and suggests works that needs to be done, either research or development.

We hypothesize that attention to these desiderata would lead to more correct local reapplications of LHS KO. The TRIPOD statement for publishing reports about predictive models gets closer to what we intend, but their focus does not address dissemination for deployment in health systems outside the locale of creation.[42] However, it is the presumption of evidence collection through rigorous research that does not obtain, necessarily, in the enthusiasm of learning health systems to share the models derived from their local data. It is just the interstices between evidence-based prediction rules and natural experiments of LHSs that these desiderata are intended to fill.

We are aware that our discussion of influence diagrams has been predicated on KOs as statistical models. That presumption may not apply in the case of machine learning or similar methods. However, the desiderata still do, and the meta-decision remains coherently modeled as a decision model. We are also aware that the local activities suggested here require workforce competencies probably beyond what local institutions host currently. Whether this lack of competence means that staff must upskill or whether consulting services are required, in either case, such skills are required for safe and effective reuse of KO. Yet, if such skills are made available, and if the research and development listed above are carried through, we can look forward to true learning among LHS as they share their knowledge.

Influence diagrams can be used in 3 ways: First, they can be the primary knowledge representation of the KO, and second, they can serve as a meta-model for the use of the KO, enabling 2 decisions: what action to take in the individual case and how to use the KO at the population level. Our treatment has addressed all of these. While our comments have been addressed primarily at those implementing and deploying knowledge objects, our concerns are predicated on the impact use of these objects has on the lives of the patients to whom these objects are applied.

ACKNOWLEDGEMENTS

Tarsha Darden and Graham Williams for bias review. Thanks to an anonymous reviewer for the suggestion of a desideratum regarding self-composition and to Patti Brennan, regarding the experience with individual artifacts. HPL: PCORI funding CDRN-1306-04912.

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

REFERENCES

1. Flynn AJ, Friedman CP, Boisvert P, Landis-Lewis Z, Lagoze C. The Knowledge Object Reference Ontology (KORO): a formalism to support management and sharing of computable biomedical knowledge for learning health systems. *Learn Heal Syst* [Internet]. 2018;2(2): e10054. https://doi.org/10.1002/lrh2.10054

2. Smith M, Saunders R, Stuckhardt L, McGinnis JM, Editors; Best care at lower cost: the path to continuously learning health care in America [Internet]. Committee on the Learning Health Care System in America; Institute of Medicine. 2012:450 p. https://www.nap.edu/read/13444/chapter/1

3. Plsek PE, Greenhalgh T. Complexity science: the challenge of complexity in health care. *BMJ*. 2001;323(7313):625-628.

4. Wallace E, Smith SM, Perera-Salazar R, et al. Framework for the impact analysis and implementation of Clinical Prediction Rules (CPRs). 2018. https://doi.org/10.1186/1472-6947-11-62. Accessed June 5, 2018.

5. The power of big data must be harnessed for medical progress. *Nature [Internet]*. 2016;539(7630):467–468. http://www.ncbi.nlm.nih.gov/pubmed/27882983. Accessed June 5, 2018.

6. Shah ND, Steyerberg EW, Kent DM. Big Data and predictive analytics. *JAMA [Internet]*. 2018;320(1):27-28. https://doi.org/10.1001/jama.2018.5602. Accessed June 5, 2018.
7. Embi PJ, Payne PRO. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. *Med care [internet]*. 2013;51(8 Suppl 3):S87-S91. http://www.ncbi.nlm.nih.gov/pubmed/23793052
8. Hripcsak G, Ludemann P, Pryor TA, Wigertz OB, Clayton PD. Rationale for the Arden syntax. *Comput Biomed Res*. 1994;27(4):291-324.
9. HL7.org. Introducing FHIR [Internet]. 2015. http://www.hl7.org/implement/standards/fhir/summary.html. Accessed October 19, 2015.
10. CDS Hooks [Internet]. 2018. http://cds-hooks.org/. Accessed February 23, 2018.
11. FORCE11. The FAIR Data Principles [Internet]. 2018. https://www.force11.org/group/fairgroup/fairprinciples. Accessed February 23, 2018.
12. Greenes RA (Ed). *Clinical Decision Support: The Road to Broad Adoption*. Second ed. Amsterdam: Elsevier; 2014.editor
13. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. *N Engl J Med*. 1980;302(20):1109-1117.
14. Djulbegovic B, Elqayam S, Reljic T, et al. How do physicians decide to treat: an empirical evaluation of the threshold model. *BMC Med Inform Decis Mak*. 2014;14(1):47.
15. Anand V, Downs SM. An empirical validation of recursive noisy OR (RNOR) rule for asthma prediction. *AMIA Annu Symp proceedings AMIA Symp* [Internet]. 2010 Nov 13 [cited 2018 Feb 26]2010;16-20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21346932
16. Matheny ME, Ohno-Machado L, Resnic FS. Discrimination and calibration of mortality risk prediction models in interventional cardiology. *J Biomed Inform*. 2005;38(5):367-375.
17. Walsh CG, Sharman K, Hripcsak G. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *J Biomed Inform* [Internet]. 2017;76:9-18. https://doi.org/10.1016/j.jbi.2017.10.008
18. Lehmann HP, Hinton R, Morello P, Santoli J. Developmental dysplasia of the hip practice guideline: technical report. Committee on Quality Improvement, and Subcommittee on Developmental Dysplasia of the Hip. *Pediatrics*. 2000;105(4):e57.
19. Centers for Disease Control (CDC). Preventing lead poisoning in young children—United States. *MMWR Morb Mortal Wkly Rep [Internet]*. 1985;34(5):66–8, 73. http://www.ncbi.nlm.nih.gov/pubmed/2982087. Accessed February 26, 2018.
20. DeBaun MR, Sox HC. Setting the optimal erythrocyte protoporphyrin screening decision threshold for lead poisoning: a decision analytic approach. *Pediatrics*. 1991;88(1):121-131.
21. CDC. New blood lead level information [Internet]. 2017. https://www.cdc.gov/nceh/lead/acclpp/blood_lead_levels.htm. Accessed February 23, 2018.
22. Centers for Disease Control and Prevention (CDC). Trends in blood lead levels among children—Boston, Massachusetts, 1994-1999. *MMWR Morb Mortal Wkly Rep [Internet]*. 2001;50(17):337-339. http://www.ncbi.nlm.nih.gov/pubmed/11465903. Accessed March 3, 2018.
23. Kennedy C, Yard E, Dignam T, Buchanan S, Condon S, Brown MJ. Blood lead levels among children aged <6 years—Flint, Michigan, 2013-2016. *Morb Mortal Wkly Rep*. 2016;65(25):650-654.
24. Gold MR, Siegel JE, Russell LB, Weinstein MC. *Cost-effectiveness in health and medicine*. New York: Oxford University; 1996.
25. von Neumann J, Morgenstern O. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press; 1947.
26. Torrance GW, Feeny D. Utilities and quality-adjusted life years. *Int J Technol assess health care [internet]*. 1989;5(4):559-575. http://www.ncbi.nlm.nih.gov/pubmed/2634630
27. Neumann PJ, Sanders GD, Russell LB, Siegel JE, Ganiats TG. *Cost effectiveness in health and medicine*. 2nd ed. New York: Oxford University; 2017.
28. Brooks RG. *The EuroQol Group after 25 years [Internet]*. Springer; 2013. http://firstsearch.oclc.org.ezp.welch.jhmi.edu/WebZ/FSFETCH?fetchtype=fullrecord:sessionid=fsapp2-34708-jecbvif7-ultz9u:entitypagenum=8:0:recno=2:resultset=2:format=FI:next=html/record.html:bad=error/badfetch.html:entitytoprecno=2:entitycurrecno=2:numrecs. Accessed March 4, 2018.
29. Feeny D, Furlong W, Torrance GW, et al. Multiattribute and single-attribute utility functions for the health utilities index mark 3 system. *Med Care*. 2002;40(2):113-128.
30. David Lunn, Chris Jackson, Nicky Best, Andrew Thomas David Spiegelhalter. *The BUGS book: a practical introduction to Bayesian analysis*. by Chapman and Hall/CRC; 2012.
31. Briss PA, Matte TD, Schwartz J, Rosenblum LS, Binder S. Screening young children for lead poisoning—guidance for state and local public health officials. Screening young children for lead poisoning: guidance for state and local public health officials [internet]. Atlanta, GA: Centers for Disease Control and Prevention; 1997. P. Appendix B4. https://www.cdc.gov/nceh/lead/publications/screening.htm
32. Howard RA. *From influence to relevance to knowledge*. 1988.
33. Smith JQ. Influence diagrams for statistical modelling. *Ann Stat*. 1989;17(2):654-672.
34. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. The Morgan Kauffman Series in Representation and Reasoning*. San Mateo, CA: Morgan Kauffmann Publishers, Inc.; 1988.
35. Owens DK, Shachter RD, Nease RF. Represention and analysis of medical decisions problems with influence diagrams. *Med Decis Making*. 1997;17(3):241-262.
36. Bowe TR. Measuring patient preferences: rating scale versus standard gamble. *Med Decis Making*. 1995;15(3):283-288.
37. Xie F, Pickard AS, Krabbe PFM, Revicki D, Viney R, Devlin N, Feeny D. A Checklist for Reporting Valuation Studies of Multi-Attribute Utility-Based Instruments (CREATE). *Pharmacoeconomics [Internet]*. 2015;33(8):867–877. https://www.ncbi.nlm.nih.gov/pubmed/26026667. Accessed February 27, 2018.
38. Savage LJ. *The Foundations of Statistics*. New York: Dover Publications; 1972.
39. Dolan JG, Boohaker E, Allison J, Imperiale TF Patients' preferences and priorities regarding colorectal cancer screening. *Med Decis Mak [Internet]*. 2013;33(1):59–70. http://www.ncbi.nlm.nih.gov/pubmed/22895558. Accessed February 27, 2018.
40. IJzerman MJ, van Til JA, Bridges JFP. A comparison of analytic hierarchy process and conjoint analysis methods in assessing treatment alternatives for stroke rehabilitation. *Patient Patient-Centered Outcomes Res [Internet]*. 2012;5(1):45-56. http://www.ncbi.nlm.nih.gov/pubmed/22185216. Accessed February 27, 2018.
41. Moreau L, Gil Y, Lebo T, Mccusker J. PROV-DM: the PROV data model W3C recommendation 30 April 2013.2014. http://www.w3.org/TR/2013/REC-prov-dm-20130430/. Accessed July 2, 2018.
42. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Eur Urol*. 2015;67(6):1142-1151.
43. Teach RL, Shortliffe EH. An analysis of physician attitudes regarding computer-based clinical consultation systems. *Comput Biomed Res*. 1981;14(6):542-558.
44. Cook DA, Teixeira MT, Heale BSE, Cimino JJ, Del Fiol G. Context-sensitive decision support (infobuttons) in electronic health records: a systematic review. *J Am Med Inform Assoc*. 2017;24(2):460-468.
45. Austrian JS, Jamin CT, Doty GR, Blecker S. Impact of an emergency department electronic sepsis surveillance system on patient mortality and length of stay. *J Am Med Informatics Assoc [Internet]*. 2017;0(December);1-7. http://academic.oup.com/jamia/article/doi/

10.1093/jamia/ocx072/4096536/Impact-of-an-emergency-department-electronic

46. Van Der Sijs H, Aarts J, Vulto A, Berg M. Overriding of drug safety alerts in computerized physician order entry. *J Am Med Informatics Assoc [Internet]*. 2006;138-148. http://www.sciencedirect.com/science/article/pii/S1067502705002483
47. Langlotz CP, Shortliffe EH. A methodology for generating computer-based explanation of decision-theoretic advice. *Med Decis Making*. 1988;8(4):290-303.
48. Gall J. *The Systems Bible*. General Systemantics: Walker, MN; 2002.
49. Constantinou AC, Yet B, Fenton N, Neil M, Marsh W. Value of information analysis for interventional and counterfactual Bayesian networks in forensic medical sciences. *Artif Intell Med [Internet]*. 2016;66:41-52. http://linkinghub.elsevier.com/retrieve/pii/S0933365715001098. Accessed December 3, 2017.
50. Osheroff JA, Teich JM, Levick D, et al. *Improving Outcomes with Clinical Decision Support: An Implementer's Guide*. Second Edition. Second ed. Health Information Management Systems Society: Chicago, IL; 2012.
51. Sonnenberg GBCFA. *Decision making in health care: theory, psychology, and applications*. 2nd ed. Cambridge, UK: Cambridge University Press; 2012.
52. Shabtai I, Leshno M, Blondheim O, Kornbluth J. The value of information for decision-making in the healthcare environment. *Stud health Technol inform [internet]*. 2007;127:91-97. http://www.ncbi.nlm.nih.gov/pubmed/17901602
53. Meltzer DO, Hoomans T, Chung JW, Basu A. Minimal modeling approaches to value of information analysis for health research. *Med Decis making [internet]*. 2011;31(6):E1-E22. http://www.ncbi.nlm.nih.gov/pubmed/21712493
54. Rothman KJ, Greenland S, Associate TLL. *Modern epidemiology*. 3rd ed. Hastings cent rep [internet]; 2014 44 Suppl 2:insidebackcover. http://www.ncbi.nlm.nih.gov/pubmed/24644503
55. Szklo M, Nieto FJ. *Epidemiology: beyond the basics [Internet]*. Jones and Bartlett Publishers; 2007. http://catdir.loc.gov/catdir/toc/fy0705/2006006230.html
56. Delgado-Rodriguez M. Bias. *J Epidemiol Community Heal [Internet]*. 2004;58(8):635-641. https://jech.bmj.com/cgi/doi/10.1136/jech.2003.008466
57. Lehmann HP, Shachter RD. In: Heckerman D, Mamdani A, eds. *End-user construction of influence diagrams for Bayesian Statistics*. Washington, DC.: Morgan Kaufmann Publishers, San Mateo, CA; 1993 p. 48–54.
58. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: evolution, critique and future directions. *Stat Med*\. 2009/07/25. 2009;28(25):3049-3067.
59. Thomas A, Lunn D, Spiegelhalter D, Best N. Package 'BRugs' [Internet]. https://cran.r-project.org/web/packages/BRugs/BRugs.pdf
60. Gelman A, Lee D, Guo J. Stan: a probabilistic programming language for Bayesian inference and optimization. *J Educ Behav Stat [Internet]*. 2015;40(5):530-543. http://jeb.sagepub.com/cgi/doi/10.3102/1076998615606113
61. Kahn MG, Callahan TJ, Barnard J, et al. A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *EGEMS* (Washington, DC). 2016;4(1):9-11.
62. McDonald CJ. Use of a computer to detect and respond to clinical events: its effect on clinician behavior. *Ann Intern Med*. 1976;84(2):162-167.
63. Cabana MD, Rand CS, Powe NR, et al. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA [Internet]*. 1999;282(15):1458-1465. http://www.ncbi.nlm.nih.gov/pubmed/10535437. Accessed December 8, 2014.
64. Grimshaw J, Eccles M, Tetroe J. Implementing clinical guidelines: current evidence and future implications. *J Contin Educ Health Prof*. 2004;24(Suppl 1):S31-S37.
65. Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: a scalable multi-trait metric for quantifying population representativeness of individual clinical studies. *J Biomed Inform [Internet]*. 2016;63:325-336. http://www.ncbi.nlm.nih.gov/pubmed/27600407. Accessed November 20, 2017.
66. Soualmia LF, Charlet J. Efficient results in semantic interoperability for health care. *IMIA Yearb [Internet]*. 2016;(1):184-187. http://www.schattauer.de/index.php?id=1214&doi=10.15265/IY-2016-051
67. Plutarch. In: Perrin B, ed. editor*Theseus [Internet]*. Cambridge, UK: Harvard University Press; 1914. http://data.perseus.org/texts/urn:cts:greekLit:tlg0007.tlg001
68. Wilkinson MD, Dumontier M, IjJ A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data [Internet]*. 2016;3:160018. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4792175&tool=pmcentrez&rendertype=abstract
69. STATO Statistics Ontology [Internet]. 2014. http://stato-ontology.org. Accessed January 30, 2018.
70. Sim I, Tu SW, Carini S, et al. The ontology of clinical research (OCRe): an informatics foundation for the science of clinical research. *J Biomed Inform*. 2014;52:78-91.
71. van Valkenhoef G, Tervonen T, de Brock B, Hillege H. *Clinical trials information in drug development and regulation: existing systems and standards*. Netherlands: University of Groningen Press; 2012.
72. McCulloch V. Maryland, (full text) : 17 U.S. 316 (1819) : Justia US Supreme Court Center [Internet]. 1819. https://supreme.justia.com/cases/federal/us/17/316/case.html. Accessed June 4, 2018.
73. Walter SD. The partial area under the summary ROC curve. *Stat Med*. 2005;24(13):2025-2040.
74. Ma H, Bandos AI, Rockette HE, Gur D. On use of partial area under the ROC curve for evaluation of diagnostic performance. *Stat Med*. 2013;32(20):3449-3458.
75. Meltzer DO, Hoomans T, Chung JW, Basu A, AHRQ. Minimal modeling approaches to value of information analysis for health research [Internet]. Methods Future Research Needs Report. Rockville (MD): Prepared by the University of Chicago Medical Center and the Blue Cross and Blue Shield Association Technology Evaluation Center Evidence-Based Practice Center Contract No. 29007-10058; 2011. www.effectivehealthcare.ahrq.gov/reports/final.cfm

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Lehmann HP, Downs SM. Desiderata for sharable computable biomedical knowledge for learning health systems. *Learn Health Sys*. 2018;2:e10065. https://doi.org/10.1002/lrh2.10065