ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 124–132

Journal of Clinical Epidemiology

VARIANCE AND DISSENT

The comprehensive diagnostic study is suggested as a design to model the diagnostic process

Norbert Donner-Banzhoff[a,*], Jörg Haasenritter[a], Eyke Hüllermeier[b], Annika Viniol[a], Stefan Bösner[a], Annette Becker[a]

[a]*Department of General Practice/Family Medicine, University of Marburg, Karl-von-Frisch Str. 4, D-35043 Marburg, Germany*
[b]*Department of Knowledge Engineering & Bioinformatics, University of Marburg, D-35043 Marburg, Germany*

Accepted 7 May 2013; Published online 28 November 2013

Abstract

Objectives: The classical diagnostic cross-sectional study has a focus on one disease only. Generalist clinicians, however, are confronted with a wide range of diagnoses. We propose the ''comprehensive diagnostic study design'' to evaluate diagnostic tests regarding more than one disease outcome.

Study Design and Setting: We present the secondary analysis of a data set obtained from patients presenting with chest pain in primary care. Participating clinicians recorded 42 items of the history and physical examination. Diagnostic outcomes were reviewed by an independent panel after 6-month follow-up (n = 710 complete cases). We used Shannon entropy as a measure of uncertainty before and after testing. Four different analytical strategies modeling specific clinical ways of reasoning were evaluated.

Results: Although the ''global entropy'' strategy reduced entropy most, it is unlikely to be of clinical use because of its complexity. ''Inductive'' and ''fixed-set'' strategies turned out to be efficient requiring a small amount of data only. The ''deductive'' procedure resulted in the smallest reduction of entropy.

Conclusion: We suggest that the comprehensive diagnostic study design is a feasible and valid option to improve our understanding of the diagnostic process. It is also promising as a justification for clinical recommendations.

Keywords: Diagnosis; Chest pain; Primary health care; Decision support techniques; Information theory; Research design

1. Introduction

1.1. Established study designs

The cross-sectional study has become the standard design to evaluate the efficacy of diagnostic tests [1,2]. For each patient in the study, the result of an index test, that is, the maneuver to be investigated, is compared with a reference standard. Although several tests are sometimes combined in the form of a clinical prediction rule (CPR), the focus in this kind of study is on one disease only. Phase III diagnostic studies include patients in whom clinicians suspect the disease under consideration to be present. These studies result in estimates of how much the likelihood of a particular disease is changed by test results.

From the investigation of a diagnostic test, we distinguish ''studies investigating symptoms.'' Patients with defined symptoms are recruited to establish the prevalence of the symptom, its etiology (underlying disease), and/or its prognosis [3]. The focus can be on a single disease or a broad spectrum of etiologies. Based on the resulting information on practice prevalence of underlying diseases, clinicians can prioritize their diagnostic workup of patients presenting with the symptom under consideration. What proportion of patients with rectal bleeding have colorectal carcinoma? What proportion of patients presenting with fatigue suffer from depression? These are examples for questions answered by this kind of study design. Although studies investigating symptoms are also cross-sectional, they serve a different purpose and report data that differ from a diagnostic accuracy study.

Results from a cross-sectional diagnostic study are particularly useful for the *late* stage of a clinician's diagnostic reasoning, when the possibilities have been narrowed to a small number of hypotheses or even only one. Study results help clinicians decide whether a test changes the likelihood of a particular hypothesis to a relevant degree. Studies investigating symptoms, however, include patients

Conflict of interest: None.

* Corresponding author. Tel.: +49-6421-2865120; fax: +49 6421 286 5121.

E-mail address: Norbert@staff.uni-marburg.de (N. Donner-Banzhoff).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.05.019

What is new?

Key findings

- We provided a novel design to retrieve and evaluate complex diagnostic data. Using a working example (study of 710 patients presenting with chest pain in primary care), we performed and compared different analytical strategies most of which emulate cognitive strategies used by clinicians.

What this add to what was known?

- Previous diagnostic study designs have considered only one disease each. The comprehensive diagnostic study (CDS) design refers to several diseases or disease categories. Moreover, a large number of diagnostic tests may be evaluated within this paradigm.

What is the implication and what should change now?

- Especially in generalist settings, clinicians process a wide array of diagnostic data. The CDS will allow the selection of valid tests and the development of clinical prediction rules with broader application than previous investigations. Moreover, data sets thus collected will improve our understanding of diagnostic uncertainty and appropriate clinical strategies.

on the basis of their (usually main) complaint, before any diagnostic effort. They can thus provide guidance regarding the general direction of the diagnostic workup.

1.2. The comprehensive diagnostic study

1.2.1. Definition

We suggest that both designs can be combined to arrive at a more thorough understanding of clinicians' diagnostic processes and improve clinical recommendations in this area. We suggest calling this study type ''comprehensive diagnostic study'' (CDS). Conducting a CDS would imply (1) recruiting patients on the basis of a particular symptom or finding, (2) recording a defined set of further symptoms, findings, and investigations as potentially valid diagnostic tests, (3) determining the final cause, that is, diagnostic outcome (disease), for each patient, and (4) analyzing results for disease probabilities occurring in the sample and how these are modified by diagnostic information.

1.2.2. Design

Within a CDS, patients with clinically relevant symptoms or findings are recruited. ''Relevant'' means that clinicians would be interested in how often certain diseases occur in these patients and how additional clinical information changes their probabilities. Patients with chest pain or vertigo/dizziness would be an example but also findings such as hepatomegaly or raised liver enzymes. Investigators should pay particular attention to their setting. With unselected patients recruited in primary care, they will obtain different results than from patients who have been referred to secondary care.

Because CDSs investigate the interplay of relevant modifiers of disease probabilities, choice of tests to be considered within the study is critical. The studies discussed here typically include items from the history and physical examination. However, further investigations can be included as long as logistically possible and ethically justifiable. Because evidence on the accuracy of the history and physical is often limited, the inclusion of tests will often be based on medical tradition, for example, textbooks or pathophysiological considerations. Surveys of clinicians can help to identify heuristics and rules that have developed in practice but have escaped scientific attention [4,5]. Ideally, all symptoms and signs that can modify the probability of one of the considered outcomes should be evaluated.

For each patient in the study, the disease associated with the symptom of entry, that is, the diagnostic outcome, has to be identified. This can result in a large number of diagnostic categories with widely differing probabilities and prognostic implications. Although this may limit the precision of analyses, investigators should try to establish each patient's diagnosis as detailed as possible. Broader categorizations can be introduced at the analysis stage.

Investigators have to decide whether and to what degree the protocol should prescribe specific tests to establish diagnostic outcomes. This often requires a large number of tests, some of which may be invasive and/or costly. Both may affect compliance of physicians and patients with study procedures if they depart too much from established routines. Alternatively, investigations can be left at the discretion of treating physicians. This, however, will lead to considerable heterogeneity of tests performed, partial verification bias, and uncertainty in outcome adjudication. Contrary to the goal of developing new diagnostic strategies, in that case, the data set will reflect the preconceptions of participating physicians. A follow-up of study patients, which has been proposed for cross-sectional diagnostic studies, will at least reveal severe chronic conditions such as cancer or coronary heart disease (CHD) [6].

For a CDS, a procedure must be described to establish the diagnostic outcome in each patient. Preferably, an independent reference panel reviews data provided by clinicians and patients. If available, specialist investigations and data obtained at follow-up visits will be taken into account. Relying on clinicians' diagnoses is a low-cost alternative, which has the disadvantage of highly variable diagnostic standards.

1.2.3. Analysis

A CDS allows modeling the uncertainty arising from a defined clinical scenario. Moreover, investigators can assess whether diagnostic tests performed within the study reduce uncertainty and to what degree. While the conventional study design estimates the probability of only one condition at a time, the analysis of CDS data has to accommodate several diseases or disease categories. Established measures of test performance such as sensitivity, specificity, or likelihood ratios are restricted to the binary case. We suggest the broader concept of Shannon's entropy to incorporate also clinical uncertainty arising from a situation with multiple diagnostic outcomes [7]. Entropy is a function of the probabilities of diagnostic outcomes considered in a defined clinical situation.

$$H(D) = -\sum P_i \log_2 P_i \quad (1)$$

In this equation, P_i is the probability of the ith disease out of a set of n mutually exclusive diseases. Entropy is measured in bits of information; $\log_2$ is the binary logarithm. Entropy increases with the number of possible disease states. It decreases as one diagnosis becomes much more likely than others. Entropy derived from information theory thus provides an intuitively valid measure of clinical uncertainty [8].

Mutual information has been suggested as an index of diagnostic test performance within this paradigm [9,10]. Once we understand the results of a diagnostic test and the disease outcome as random variables, mutual information quantifies the amount of information that the test provides about the diseases under consideration (for a more detailed definition, see the subsection Data analysis). A decision can then be made whether diagnostic testing reduces uncertainty (entropy).

Data sets derived from CDS can be large and complex. Theoretically, the whole array of statistical (machine) learning can be applied [11]. However, the main purpose of a CDS is not a computer program (expert system) because this is unlikely to be used in everyday practice. It should rather aim at recommendations regarding tests to be used in defined clinical situations. The order or combination (CPRs) of test results can be investigated in this context. Which tests can be omitted because they do not reduce clinical uncertainty to a sufficient degree is also of interest. To achieve this, we suggest the following analytical strategies which partly emulate cognitive strategies used by clinicians:

1.2.3.1. Global entropy. All recorded tests are examined sequentially regarding their ability to reduce the overall entropy. The latter is measured by mutual information [9,10]. According to the result of the test that reduces entropy most, the sample is split into two groups (binary case). On each of these, the remaining tests are evaluated regarding their ability to reduce entropy in the sample. Again, two groups ensue according at each branch, which are again investigated in a similar way. This procedure often leads to a complex decision tree with a large number of branches and decision chains.

1.2.3.2. Deductive (hypothesis based). Considering their frequency, severity, and treatability, investigators establish a priority list of possible diseases expected to occur in the sample. Using the CDS data set, each is evaluated in turn by symptoms and signs modifying their likelihood. Once a patient scores positive for a diagnosis, search is stopped and treatment is begun. Depending on the setting, this may also include specialist referral for further workup. Cases scoring below the threshold will remain in the pool of undiagnosed subjects and be evaluated for the next hypothesis and so on. This kind of analysis corresponds to the hypothetico-deductive clinical method. According to this theory, very early in the encounter with the patient, hypotheses (possible diseases) ''pop'' into the clinician's mind and guide further data collection [12].

Ideally, simple scores to evaluate diagnostic hypotheses derived from other studies are used for this kind of analysis. Alternatively, a decision rule is derived from the data set. Cases scoring above a defined treatment threshold are regarded as having that particular disease.

1.2.3.3. Inductive (test based). Sometimes, a single symptom or sign by being positive increases the likelihood of a particular disease to such a degree that the diagnostic process can be stopped at this point (pathognomonic sign). Although this occurs only rarely, clinicians paying attention to this possibility increase the efficiency of their diagnostic reasoning considerably. This strategy always has to be combined with other approaches to be applied to the remaining majority of patients negative for pathognomonic signs.

1.2.3.4. Fixed set. This strategy implies the identification of a set of items that are evaluated on every patient irrespective of particular findings. The items are chosen according to their ability to reduce the entropy in the study data set. This strategy is analogous to the review of systems as part of a full medical history. Here, a defined set of questions referring to one organ system is asked, as part of the hospital admission procedure or with frequent presentations.

2. Methods

2.1. Data set

To illustrate the design of a CDS and related analytical approaches, we performed a secondary analysis of the first Marburg chest pain study. Design and conduct of this study have been described elsewhere [13]. Briefly, 74 general practitioners (GPs) recruited consecutively patients aged

≥35 years who presented with chest pain as the primary or secondary complaint. GPs took a standardized history and performed a physical examination. Results of 42 single diagnostic tests were thus obtained. These included pain quality; location and severity; modifying factors, for example, chest pain associated with food; associated symptoms, for example, shortness of breath; and known risk factors, for example, diabetes or known CHD. Patients and GPs were contacted 6 weeks and 6 months after the index consultation. All available information about the course of chest pain, treatments including hospitalizations and drugs, and diagnostic procedures initiated by GPs or specialists was retrieved. An independent expert panel of one cardiologist, one GP, and one research staff member reviewed each patient's data and established the reference diagnosis by deciding which disease was the underlying cause for the chest pain at the time of index consultation. For the analysis presented here, we collapsed 28 reference diagnoses into nine larger categories (Table 1). In 710 of 1,212 patients, information on all 42 symptoms and signs was available. Analyses presented in this article are based on this sample.

2.2. Data analysis

We calculated entropy $H(D)$ as a measure of our uncertainty about the disease state or diagnostic outcome using formula (1). $H(D|T)$ quantifies our average uncertainty about the disease state (D) given the results of a test (T) and is calculated by

$$H(D|T) = \sum_{i=1}^{m} P(t_i) \times H(D|T = t_i), \qquad (2)$$

where $P(t_i)$ is the probability that a patient has the test result i and $H(D|T = t_i)$ is the entropy about the disease status within the sample of patients with the test result i [10].

The difference between $H(D)$ and $H(D|T)$ is the mutual information $I(D;T)$. It quantifies the average amount of information gained by performing a diagnostic test, a large I stands for a test achieving a large reduction of uncertainty [10].

All calculation were performed within R, 2.14.1 (R Foundation for Statistical Computing, Vienna, Austria) (http://www.R-project.org) using the entropy 1.1.7 package [14], and Rweka 3.7.5, an R interface to Weka [15,16].

Table 1. Diagnostic categories/outcomes considered in the analysis

Diagnostic category	Explanation
Stable coronary heart disease	
Acute coronary syndrome	Myocardial infarction, unstable angina
Nonischemic cardiovascular disorders	Hypertension, heart failure, cardiac arrhythmia, myocarditis, heart valve defect, pulmonary embolism
Severe respiratory disorders	Chronic obstructive pulmonary disease, pneumonia,
Benign respiratory disorders	Cold, bronchitis
Benign digestive disorders	Gastroesophageal reflux disease, gastritis
Psychogenic causes	Depression, anxiety, stress-related and somatoform disorders
Chest wall syndrome	Various musculoskeletal reasons and disorders
Others	No specific diagnosis possible

2.2.1. Individual index tests and fixed set

Before evaluating complex strategies, we calculated the mutual information of each item (test) individually and ordered them according to their ability to reduce the entropy. We then calculated the mutual information of a set of the best two items, of the best three items, and so on. We plotted the post-test entropy $H(D|T)$ of the best one, two, three, four, and five fixed test sets, respectively. This plot can help to decide whether the expense of adding another item to the set results in a relevant reduction of uncertainty. We used the best set of three tests to illustrate possible clinical implications.

2.2.2. Global entropy

Following the "global entropy" approach, we constructed a decision tree using the C4.5 algorithm. C4.5 first grows a tree using information gain as the split criterion [16]. The test/symptom or sign that minimizes the entropy of the resulting subsets is chosen to partition the data set. Information gain equals the expected value of the mutual information of a test. To avoid overfitting and unnecessary complexity, the initial tree is pruned, that is, unreliable parts of the tree are discarded. The pruning procedure is based on the comparison of the error estimates of an internal node and the nodes below it. According to the result, a subtree is replaced by a terminal node, raised or kept. Because the same data set that was used to raise the tree was used to prune the tree, the upper limits of the confidence intervals of the error rates are compared instead of their point estimates [16]. We ran C4.5 using a confidence factor, $c = 0.05$, which corresponded to a confidence level = 0.95%. In C4.5, all errors are treated as equal irrespective of the diagnostic outcome. To determine the performance of the decision tree, we calculated the accuracy that equals the number of correctly classified cases divided by the total number of cases. Furthermore, we calculated the average pretest entropy about the disease state, $H(D)$, the entropy about the disease state given the result of the classification using the tree, $H(D|T)$, and the mutual information, $I(D;\ T)$, as a measure of the diagnostic performance.

2.2.3. Deductive (hypothesis based)

We selected three conditions to be evaluated: (1) need for urgent hospital admission, (2) CHD, and (3) chest wall syndrome (CWS). For each of these, a CPR is available (Table 2). All CPRs have been derived from the current data set, they provide binary thresholds (condition yes/no). CPRs were successively applied starting with the

Table 2. Clinical predicting rules used in the deductive approach

CPR	Urgent admission score [21]	Marburg Heart Score [22]	CWS score [23]
Items and scoring	Known clinical vascular disease (1 point) History of heart failure (1 point) Home visit required (1 point) Pain not reproducible by palpation (1 point) Pain radiation to left arm (1 point)	Age/gender (female ≥65 yr, male ≥55 yr) (1 point) Known clinical vascular disease (1 point) Patient assumes cardiac origin of pain (1 point) Pain worse with exercise (1 point) Pain not reproducible by palpation (1 point)	Pain reproducible by palpation (1 point) Stinging pain (1 point) Localized muscle tension (1 point) Absence of cough (1 point)
Threshold	Test negative: 0–2 points Test positive: 3–5 points	Test negative: 0–2 points Test positive: 3–5 points	Test negative: 0–3 points Test positive: 4 points

Abbreviations: CPR, clinical prediction rule; CWS, chest wall syndrome.

"urgent hospitalization score." Cases scoring positive were selected for treatment, which in this case denoted urgent referral to hospital. The remaining cases were evaluated for CHD and so on. We calculated $H(D|T)$ and the mutual information for the whole test sequence. Additionally, we plotted the probabilities of the disease outcomes for the respective subsamples.

2.2.4. Inductive (test based)

To identify "pathognomonic" symptoms or signs, we calculated the likelihood ratios (LR) for each test and diagnostic outcome category. For each combination, we inserted the LR values >1 into a matrix with columns representing the diagnostic outcomes (diseases) and rows representing tests. In this LR matrix, the pattern indicating a pathognomonic test would be a test with a high LR for one diagnostic category, whereas LRs for the other diagnostic categories were substantially lower. We combined the tests identified in this manner in a test sequence that was successively applied to the cases in the data set. Again, we calculated $H(D|T)$ and the mutual information for the whole test sequence and we plotted the probabilities of the disease outcomes for the respective subsamples.

3. Results

3.1. Fixed set

The mutual information of the single tests ranged from 0.06 (dull chest pain) to 0.23 bits (cough) (see Appendix at www.jclinepi.com). Fig. 1 shows the effect of test combinations; the higher the number of items in the set, the higher was the decrease in the entropy.

The best three-item test set was "cough/pain reproducible by palpation/known CHD." These correspond to a combined test with eight possible outcomes. As can be seen from Fig. 2, the outcomes are clinically plausible with largely differing disease probabilities. The number of patients and entropy in the respective outcome categories ranged from 3 and 0.0 bits to 273 and 2.49 bits, respectively. Applying this combined test reduced the entropy of the data set from $H(D) = 2.35$ to $H(D|T) = 1.88$ bits. Mutual information was 0.47 bits.

The best five-item test set was "cough/pain reproducible by palpation/stabbing pain/history of hypertension/patient is anxious" with a mutual information of 0.69 bits.

3.2. Global entropy

The final decision tree exploited the information of 19 of the 42 tests included in the data set and resulted in 24 decision/internal and 25 terminal nodes (see Appendix at www.jclinepi.com). The accuracy was 65.8%. The entropy and the number of cases in the terminal nodes ranged from 0.0 to 2.3 bits and from 2 to 215, respectively. Applying this decision rule reduced the average entropy from $H(D) = 2.35$ to $H(D|T) = 1.65$ bits. The mutual information was 0.7 bits.

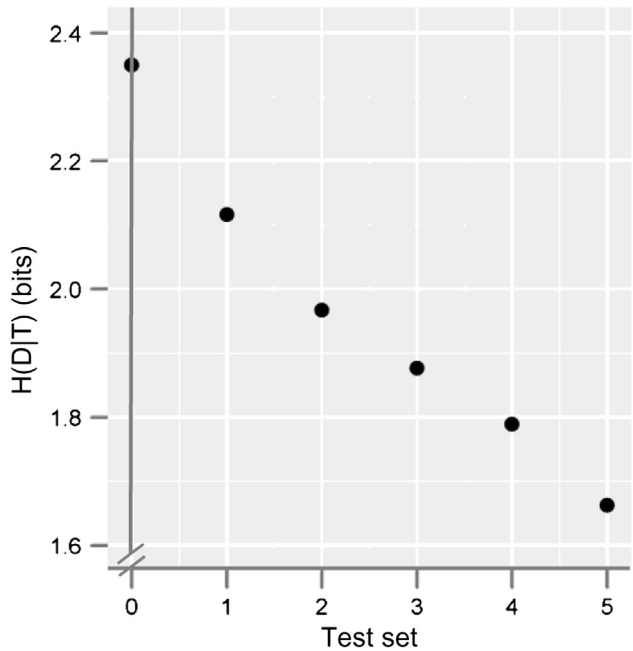

Fig. 1. Maximum decrease in entropy using optimal sets of one, two, three, four, and five tests. Y-axis shows the entropy, given the information of the respective tests. X-axis shows the number of items in the test set and names the respective optimal test set. 0: pretest entropy, no test applied; 1: cough; 2: cough and pain reproducible by palpation; 3: cough, pain reproducible by palpation, and known CHD; 4: cough, pain reproducible by palpation, known CHD, and stabbing pain; 5: cough, pain reproducible by palpation, stabbing pain, history of hypertension, and patient is anxious. CHD, coronary heart disease.

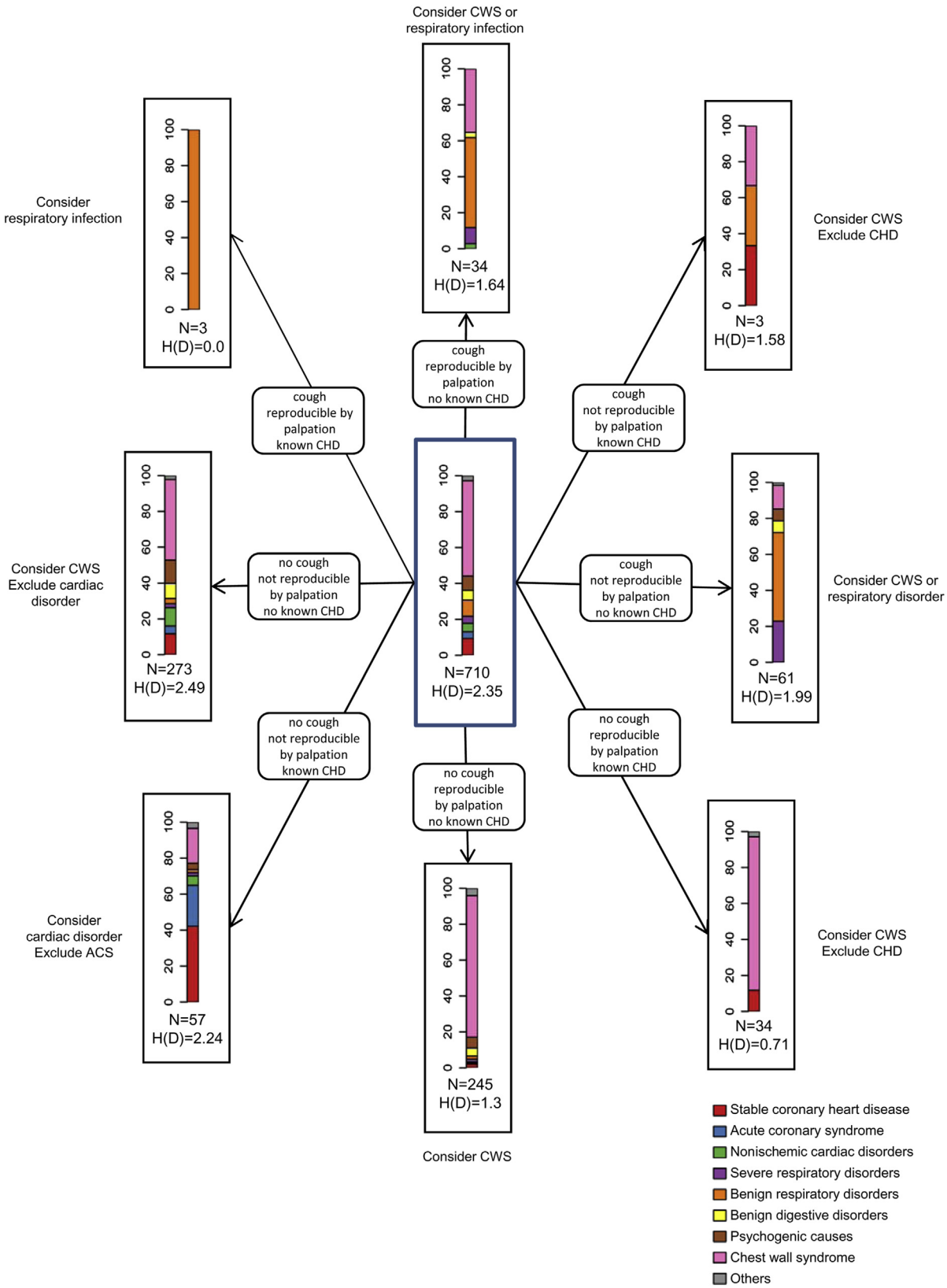

Fig. 2. Result of the optimal (maximal decrease in entropy) set of three tests: cough, pain reproducible by palpation, and history of CHD. For each possible combination of the results of the three tests, the number of patients, the entropy, and the proportion of underlying diseases are shown. Implications for the clinical management for the resulting subsets are also suggested. Scale of the bar charts is in percentage. CWS, chest wall syndrome; CHD, coronary heart disease; *H*(*D*), entropy about disease status in bits; *N*, number of patients in the sample; ACS, acute coronary syndrome.

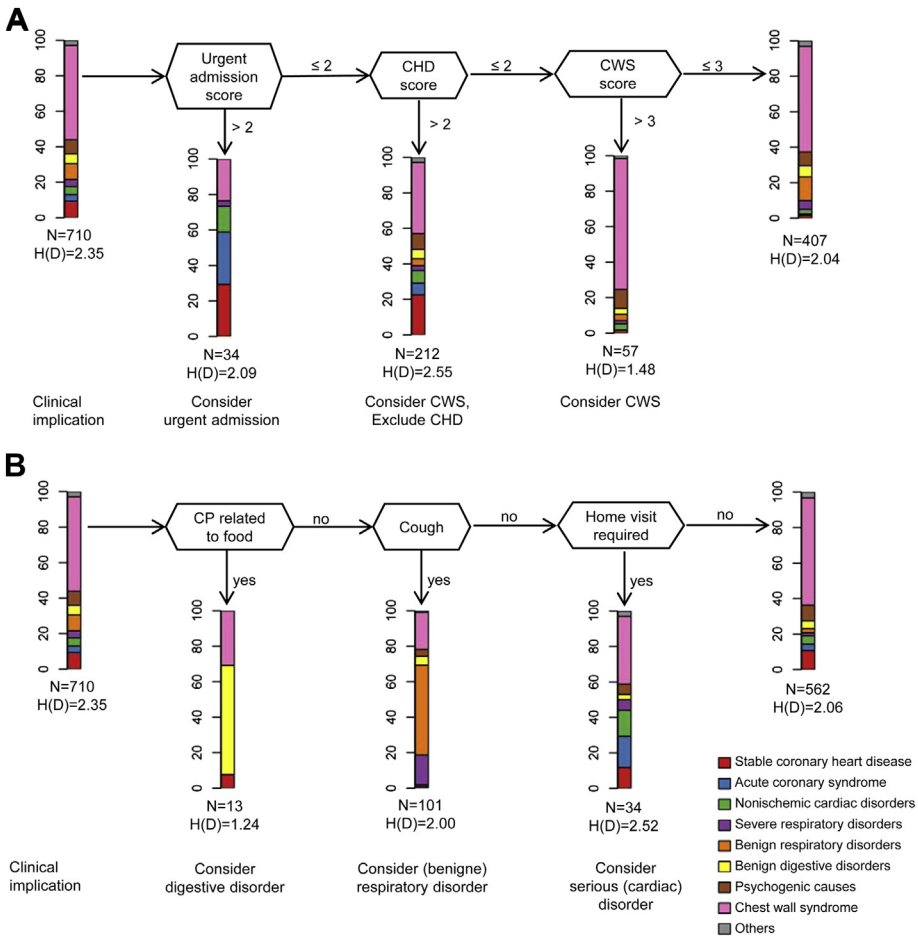

Fig. 3. (A) Results of the deductive hypothesis-based strategy. Three clinical hypotheses (''Need for an urgent admission?'' ''Has the patient an underlying CHD?'' ''Has the patient an underlying CWS?'') were successively tested by applying the respective clinical decision rule. For each resulting subset, the number of patients, the entropy, and the proportion of underlying diseases are shown. Implications for the clinical management for the resulting subsets are also suggested. (B) Results of the inductive test-based strategy. The three items/tests identified in the inductive approach as pathognomonic symptoms (''pain related to food,'' ''cough,'' and ''home visit required'') were successively applied. For each resulting subset, the number of patients, the entropy, and the proportion of underlying diseases are shown. Implications for the clinical management for the resulting subsets are also suggested. Scale of the bar charts is in percentage. CHD, coronary heart disease; CWS, chest wall syndrome; *CP, chest pain; H(D)*, entropy about disease status in bits; *N*, number of patients in the sample.

3.3. Deductive (hypothesis based)

The deductive approach resulted in four subsets of patients (Fig. 3A). In those with a positive urgent hospital admission score ($n = 34$), only five different diagnostic outcomes were left and entropy was reduced from $H(D) = 2.35$ to $H(D|T) = 2.09$ bits. For cases scoring positive on the Marburg Heart Score ($n = 212$), entropy increased slightly from 2.35 to 2.55 bits, whereas the number of diagnostic outcomes (diseases) was identical to the whole sample. In the subsample with a positive CWS score ($n = 57$), the number of diagnoses was only slightly decreased to eight, whereas the entropy was decreased to 1.48 bits. In the remaining subset, including 407 patients, the entropy was 2.04 bits. On the whole, applying this sequence of three CPRs reduced the entropy from $H(D) = 2.35$ to $H(D|T) = 2.15$ bits. The mutual information was 0.2 bits.

3.4. Inductive (test based)

Following the inductive analytical approach, we identified three tests that we considered as pathognomonic symptoms or signs (see Appendix at www.jclinepi.com for the likelihood ratio matrix): chest pain related to food (LR = 27.5 for benign digestive disorder), cough (LR = 10.3 for benign respiratory disorder), and home visit needed [LR = 4.9 for acute coronary syndrome (ACS)]. Using these tests in a sequence resulted in four subsamples (Fig. 3B). For the first subsample of patients with a high likelihood of a benign digestive disorder ($n = 13$), entropy was reduced to 1.24 bits. Cough as an additional symptom was the criterion for the second subsample ($n = 101$). The likelihood of benign respiratory infection thus increased to 50.5%. The entropy in this subsample was reduced to 2.05 bits. For patients in the third subsample, a home visit was required ($n = 34$), 44% of these had a cardiac

condition, and entropy was 2.52 bits. For the remaining patients ($n = 562$), the entropy was relatively low with 2.06 bits mainly because of a high prevalence of CWS. Overall, applying this test sequence reduced the entropy from $H(D) = 2.35$ to $H(D|T) = 2.07$. Mutual information was 0.28 bits.

4. Discussion

In a study of 710 patients presenting with chest pain in primary care practice, 42 index tests and 9 diagnostic outcome categories were evaluated. Analytical strategies resulted in widely differing reductions of uncertainty quantified as entropy.

4.1. Strategies compared

"Global entropy" analysis resulted in the largest decrease of entropy compared with the other approaches. This strategy made use of a large part of the information contained in the data set without leaving patients unassigned to a diagnostic outcome category. However, calculation of the related decision tree is very complex and requires sophisticated software. This analysis provides a reference standard for the remaining strategies, but resulting recommendations are unlikely to be used in practice because of their complexity.

The deductive (hypothesis-based) strategy used only information derived from eight index tests; three of these were part of more than one CPR. This analysis left 407 of 710 cases unassigned and achieved the least decrease in entropy compared with other analytical strategies.

The inductive (test-based) strategy made use of only three test criteria; a pool of 526 cases remained without diagnostic assignment. Chest pain being associated with food allowed the rapid selection of cases with benign digestive disease, as happened with cough (benign respiratory disease) and home visits (CHD). The overall entropy was reduced to 2.07 bits.

Using the fixed-set strategy, it was possible to name the set of five items that provided the highest reduction in entropy. However, the higher the number of items in the set, the higher is the number of possible outcomes in the combined test. Given that all items are dichotomous, the number of possible outcomes is 2^n, in which n is the number of items in the set. This can rapidly lead to a number that is not cognitively manageable any more. There is thus a trade-off between reduction in entropy by a defined set of tests and cognitive demand on the clinician.

4.2. Clinical implications

The outlined analytical strategies correspond to cognitive strategies used by clinicians when confronted with a patient. The deductive or hypothesis-based analysis reflects the hypothetico-deductive method [12]. The inductive or test-based method corresponds to "inductive foraging," which has unusual phenomena (symptoms, pattern failure, and sense of alarm) as its starting point, be that from the perspective of the patient or the physician. This strategy has been suggested as an efficient way of data collection in low-prevalence generalist areas [17]. Finally, the fixed set of routine questions (tests) parallels the review of systems. Among theories of clinical reasoning, structural semantic theory proposed by Bordage [18] corresponds to this kind of analysis. In this view, disease concepts are represented in the clinician's mind as prototypes, that is, combinations of binary tests, so-called semantic qualifiers. The data presented here suggest that the hypothetico-deductive method often suggested as the normative and descriptive standard of diagnostic reasoning [12] may have a more limited role in generalist settings. Here, other strategies such as inductive (test-based) ones should also be used.

The study design presented here will be particularly fruitful in generalist settings such as primary care practice or hospital emergency departments. In these settings, the history and physical examination comprise a large number of cheap and quick-to-conduct diagnostic tests that can contribute to efficiently triage patients with a wide range of possible disease states. However, this kind of study may also be relevant for specialized settings and findings arising from physical examination, biochemical tests, or imaging procedures.

The analyses presented generally show that uncertainty can only be reduced to a certain degree. In our view, they give a realistic picture of diagnostic uncertainty in primary care. Further investigations such as imaging or biochemical tests are of use only in relatively small subsets of patients with a high probability of specific diseases. For most patients, strategies such as "watchful waiting" maybe more appropriate. In practice, clinicians will usually combine the strategies evaluated in this article.

The global entropy strategy results in a highly complex decision tree, which reflects the overall uncertainty of the diagnostic situation but is cognitively intractable for human beings. However, an expert system based on this algorithm might be of use in clinical teaching. It could give learners immediate feedback on the validity and efficiency of their diagnostic data gathering, that is, history taking, physical examination, or investigations. As opposed to classical teaching based on predictive values or likelihood ratios, this would refer to all relevant diseases and not just one.

4.3. Limitations

The study, which our analysis is based on, was not originally designed to evaluate all possible diagnostic outcomes in patients with chest pain. Because its focus was on CHD, assignment of non-CHD diagnoses by the reference panel was often based on sparse data; some degree of misclassification is therefore possible. In some instances, reference diagnoses had to be based on the history, which might have resulted in incorporation bias. Because of these

shortcomings, we regard our analysis as illustrative for the CDS design but not suitable for clinical recommendations per se.

In the original study sample, 71 different reference diagnoses were made. To ease analysis and understanding of results, we collapsed them into nine broad categories. For these, we considered not only anatomy and physiology but also prognosis and management. In other words, we avoided grouping patients with different prognosis (benign self-limiting vs. life-threatening conditions) together in one category.

In most of our analyses, we have treated diagnostic outcome categories and related errors equally. Only in the deductive (hypothesis-based) approach, clinical priorities are reflected in the order of hypotheses evaluated. Missing life-threatening diseases such as ACS would cause more regret or chagrin in the clinician than missing benign conditions such as CWS [19,20]. Adjustment for differential weights of disease entities will become important in studies with multiple diagnostic outcomes.

The large number of variables analyzed in this kind of study results in a high degree of variability. Strategies to establish the precision of estimates, evaluate sampling error, and calculate sample size will have to be developed. As with any CPR, validation by an independent data set will have to establish the robustness of findings. We chose uncertainty (entropy) to evaluate the effect of diagnostic tests, not only for its mathematical properties but also as an intuitive measure of clinical uncertainty. However, other measures can be considered for this purpose. The same applies to other classes of statistical models than the ones used for this analysis.

4.4. Conclusion for future studies

We would like to encourage researchers to extend their diagnostic study design toward all diagnostic outcomes relevant to a particular clinical situation. Moreover, they should consider the evaluation of a large number of index tests to model diagnostic reasoning in a more comprehensive way than with the classical one-test study design. We hope that the design features and analytical strategies presented in our article will provide useful guidance for this kind of work.

Acknowledgments

The authors thank Bill Benish (Cleveland, Ohio, USA) for detailed feedback on a previous version of the manuscript.

Appendix

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2013.05.019.

References

[1] Guyatt G, Sackett DL, Haynes RB. Evaluating diagnostic tests. In: Haynes RB, Sackett DL, Guyatt GH, Tugwell P, editors. Clinical epidemiology: how to do clinical practice research. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2005:273−322.
[2] Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002;324:539−41.
[3] Donner-Banzhoff N, Kunz R, Rosser W. Studies of symptoms in primary care. Fam Pract 2001;18:33−8.
[4] Abu Hani MA, Keller H, Vandenesch J, Sonnichsen AC, Griffiths F, Donner-Banzhoff N. Different from what the textbooks say: how GPs diagnose coronary heart disease. Fam Pract 2007;24:622−7.
[5] Andre M, Borgquist L, Foldevi M, Molstad S. Asking for 'rules of thumb': a way to discover tacit knowledge in general practice. Fam Pract 2002;19:617−22.
[6] Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. J Clin Epidemiol 2003;56:1118−28.
[7] Shannon CE, Weaver W. The mathematical theory of communication. Urbana, IL: University of Illinois Press; 1998 (reprint).
[8] Benish WA. The application of information theory to diagnostic testing: a primer. Available at http://www.mutualinformation-medicine.com/. Accessed October 9, 2013.
[9] Cover TM, Thomas JA. Elements of information theory. Hoboken, NJ: Wiley-Interscience; 2006.
[10] Benish WA. Mutual information as an index of diagnostic test performance. Methods Inf Med 2003;42:260−4.
[11] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009.
[12] Elstein AS, Schulman LS, Sprafka SA. Medical problem-solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978:1978:xvi + 330 pp.
[13] Bösner S, Becker A, Haasenritter J, Abu Hani M, Keller H, Sönnichsen AC, et al. Chest pain in primary care: epidemiology and pre-work-up probabilities. Eur J Gen Pract 2009;15:141−6.
[14] Hausser J, Strimmer K. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res 2009;10:1469−84.
[15] Hornik K, Buchta C, Zeileis A. Open-source machine learning: R meets Weka. Comput Stat 2009;24:225−32.
[16] Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington, MA: Morgan Kaufmann; 2011.
[17] Donner-Banzhoff N, Hertwig R. Inductive foraging. Tactics to improve the diagnostic yield of the consultation. Eur J Gen Pract 2013[Epub ahead of print].
[18] Bordage G. Prototypes and semantic qualifiers: from past to present. Med Educ 2007;41:1117−21.
[19] Feinstein AR. The chagrin factor and qualitative decision-analysis. Arch Intern Med 1985;145:1257−9.
[20] Djulbegovic B, Tsalatsanis A, Hozo I, Vickers A. A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making. BMC Med Inform Decis Mak 2010;10:51.
[21] Bösner S, Becker A, Abu Hani M, Keller H, Sönnichsen AC, Haasenritter J, et al. Accuracy of symptoms and signs for coronary heart disease assessed in primary care. Br J Gen Pract 2010;60:246−57.
[22] Bösner S, Haasenritter J, Becker A, Karatolios K, Vaucher P, Gencer B, et al. Ruling out coronary artery disease in primary care: development and validation of a simple prediction rule. CMAJ 2010;182:1295−300.
[23] Bösner S, Becker A, Hani MA, Keller H, Sönnichsen AC, Karatolios K, et al. Chest wall syndrome in primary care patients with chest pain: presentation, associated features and diagnosis. Fam Pract 2010;27(4):363−9.