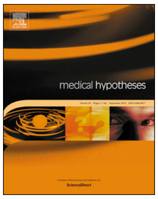
Medical Hypotheses 123 (2019) 98–100

Contents lists available at ScienceDirect

Medical Hypotheses

journal homepage: www.elsevier.com/locate/mehy



The Perception Approach: Level of Observer Variability Rating and Fusion Diagnoses

Benedikt Martin*, Bruno Märkl

Institute of Pathology, Klinikum Augsburg, Augsburg, Germany

ABSTRACT

Diagnostic uncertainty is a problem faced daily in clinical medicine. This uncertainty has many guises, such as vague wording in reports or pseudo-exact statements. High rates of intra- and interobserver variability and a corresponding low level of reproducibility are, unfortunately, issues that have to be taken into consideration in many medical fields. We hypothesise that the level of observer variability can be predicted in specific situations during the on-going assessment/diagnostic process and that the application of a communication concept based on Fusion Diagnoses might lead to improved patient treatment. A Fusion Diagnosis consists of the diagnosis itself combined with the estimated Level of Observer Variability (LOV) of the diagnostic process which is evaluated by the individual decision-maker during the on-going assessment process stated in parenthesis. Currently, the "Perceived Difficulty" of the investigator in a three-tiered grading system might be the most promising approach for the operationalisation of the LOV in the diagnostic process (D1 = easy to assess, D2 = average, D3 = difficult to assess). An example of a diagnosis according to this concept could be: Dysplastic Nevus (D3). Based on this concept, standardized clinical decision pathways could be investigated and developed. Studies would have to provide data about rational decisions in relation to the estimated LOV. For example, data might reveal that "ST-elevation myocardial infarction (D3)" as well as "no ECG abnormalities (D3)" are necessary indications for an immediate consultation of a colleague in the setting of suspected myocardial ischaemia. The concept is hypothetical and critical aspects, like over- and underestimation and the calibration of the physician, are likely limiting the benefit. Nevertheless, we assume that the suggested pragmatic concept might lead to superior patient treatment in specific fields.

Introduction

In the practice of medicine, making a decision and consequently communicating the decision are of outstanding importance. These two different abilities have to be considered as key qualifications. They are needed for the evaluation of symptoms as well as for the evaluation of numerous available medical tests and in making final diagnosis. High quality medical care is often the result of a well-communicating team of specialists. The establishment of tumor boards is an example of the interest of the medical community and its efforts in further developing effective as well as structured interdisciplinary communication. In clinical medicine diagnostic uncertainty is a problem faced daily. There are excellent reviews addressing the topic of uncertainty [1,2]. Uncertainty has many guises such as vague formulation in reports or – even more dangerous - pseudo exact statements. Phrases are not standardized which can lead to more uncertainty due to the need for interpretation thereof [3,4]. Physicians, that are not part of a certain discipline but also specialists within the field are often unaware of the limit of reproducibility of the tests and diagnoses. Pathology is considered as one of the most accurate disciplines in clinical medicine and the results are often regarded as "gold standard". Nevertheless, the diagnostic interpretations of cervical specimens and melanocytic lesions, for instance, are as just two examples where uncertainty still prevails in this field. In a study about the interpretative variability of cervical cytologic and histologic specimens the authors conclude that "Given the degree of irreproducibility that exists among well-trained pathologists, realistic performance expectations should guide use of their interpretations" [5]. Another well-designed study illustrates that "diagnoses spanning moderately dysplastic nevi to early stage invasive melanoma were neither reproducible nor accurate" [6]. The intraobserver reproducibility (IR) of cases interpreted as class II (e.g. moderate atypia, IR: 35.2%), class III (e.g., severe atypia or melanoma in situ, IR: 59.5%) and class IV (e.g., pathologic stage T1a, early invasive melanoma, IR: 63.2%) [6] appear rather as product of chance than as an established "gold standard". To add an example from cardiology, ECG interpretation errors of major importance happen in 4 to 33%, and inappropriate management because of ECG interpretations happens in up to 11% of cases [7].

Our group addressed the topic of the interobserver variability in the assessment of tumor budding in colon cancer and observed that the interobserver variability can be estimated by the investigator during the on-going assessment process [8,9]. On average, the relative risk of disagreement was more than six times higher when a case was estimated to have a rather high interobserver variability (absolute risk of disagreement for cases with estimated low interobserver variability 6.6% and absolute risk of disagreement of cases with high interobserver

* Corresponding author at: Institut für Pathologie, Klinikum Augsburg, Stenglinstraße 2, 86156 Augsburg, Germany.
E-mail address: benedikt.martin@klinikum-augsburg.de (B. Martin).

https://doi.org/10.1016/j.mehy.2019.01.002
Received 21 October 2018; Accepted 6 January 2019

variability 41.6%) [8].

This observation inspired us to propose a pragmatic approach according to diagnostic uncertainty in the form of observer variability. Physicians should be encouraged to consider uncertainty consciously in routine diagnostics and communicate it accordingly. Furthermore, physicians should be guided to appropriate decisions.

Level of Observer Variability Rating and Fusion Diagnoses

We propose a concept in which the estimated Level of Observer Variability (LOV) becomes an indicator for the quality of the diagnosis and gets communicated in a standardized approach. The Perceived Difficulty of the investigator (for example D1 = easy to assess, D2 = average, D3 = difficult to assess) could be used although it has to be elucidated which is the best operationalized parameter for the LOV. Thereby, multiple patient-related criteria become transformed during the on-going decision process and are melted together with the experience of the decision-maker in one parameter.

A Fusion Diagnosis consists of the diagnosis and, in parenthesis, of the estimated Level of Observer Variability of the diagnostic process which is evaluated by the individual decision-maker during the ongoing assessment process.

For example: ST-elevation myocardial infarction (D1) or ST-elevation myocardial infarction (D3).

The investigator provides a definite diagnosis and integrates additional information to support an appropriate interpretation by other parties. The estimation of the LOV should lead to a risk stratification concerning the likelihood of the correctness of the diagnosis. In any subsequent communication of the diagnosis, third parties would invariably be confronted with this new surrogate parameter for the quality of the diagnosis. Thereby, the interpreter of the test result or the diagnosis in question, as well as the ultimate decision-maker, would be forced to consider the certainty-level of the test results and diagnosis when deliberating on further measures.

We assume that the use of a three-tiered grading system is the most promising approach for grading the LOV given its simple and intuitive nature. However, the question of the best rating scale (for example three-tiered) requires further investigation in studies. In our tumor budding project we investigated also the parameter of the estimated Interobserver Variability, which was significantly correlated with the perceived difficulty [8].

Fields of application

We see a potential worthwhile application in all medical fields, but therein strictly restricted to fields with clinical relevance and predictable interobserver variability/uncertainty. Returning to our example of classification of melanocytic lesions, a likely result for D3 situations will be that mandatory consultation will be required given the high rates of disagreement. In routine diagnostics pathologists show difficult cases to their colleagues and clinical physicians ask colleagues for advice. But there is no good standardized procedure regarding in what cases colleagues should be consulted and at what stage/when. We suppose that criteria implemented based on the model outlined in this paper, could be advantageous because the decision-maker can use the grade of LOV as guidance. Studies could provide data about rational decisions in dependence on the estimated Level of Observer Variability.

A parameter of the decision-maker

The practice of evidence based medicine can be understood as the integration of individual clinical expertise with the best possible external evidence from systematic research [10]. While external evidence from systemic research is well defined, there is, to our knowledge, no parameter in the practice of routine diagnostic for the individual clinical expertise which is measured through. Therefore, clinical expertise can only be integrated into guidelines and recommendations in general. In this context we propose the LOV also as a surrogate parameter for the decision-maker. The LOV is also a surrogate parameter for the individual clinical expertise of the decision-maker for the performed decision-making process in this case and at this moment. On this basis, a part of the clinical expertise becomes operationalized and procedure pathways could be developed and integrated into guidelines (Fig. 1). In contrast to the threshold approach [11], the decision would be dependent on the perception of the diagnostic process instead of the estimation of the probability. We assume that on one hand this is associated with a higher degree of inaccuracy but on the other hand it is very simple and applicable.

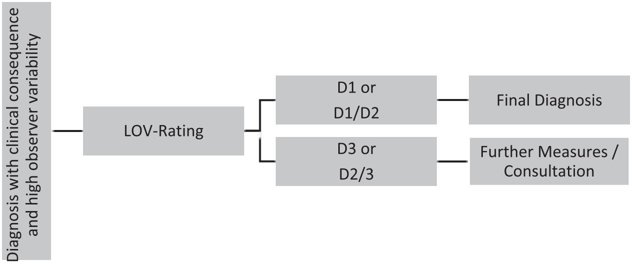

Diagnosis with clinical consequence and high observer variability
LOV-Rating
D1 or D1/D2
Final Diagnosis
D3 or D2/3
Further Measures / Consultation

Fig. 1. Illustration of a potential clinical decision pathway. D1 = easy to assess, D2 = average, D3 = difficult to assess.

The concept is simple and highly flexible because it is not dependent on quantitative or qualitative criteria, which would need to be defined in advance. It is on the one hand not as standardized as patient-related qualitative and quantitative criteria (for example length in cm or concentrations in mg/ml) but on the other hand not as unspecific as general recommendations (ovarian borderline tumors have to be consulted). The concept is located in between. In the last few decades evidence-based medicine aimed, for good reasons, to reach the highest-level of standardization of medical practice by defining patient-related criteria. In contrast, we assume that this concept offers the chance of an additional risk stratification by giving up predefined patient related criteria. Because of that, the topic of reproducibility has to be addressed thoroughly in studies evaluating the concept. Furthermore, future studies would have to evaluate the influence on the different levels of the diagnostic process and interdisciplinary communication. Issues that need to be examined critically are whether there are measurable effects concerning accuracy, number of investigations and costs and whether the implementation of LOV does alter the reputation of a diagnostician.

Critical aspects

There are several aspects which have to be critically discussed. First of all, it has to be assumed that the perceptions will suffer from over- as well as underestimation, independently which parameter is used for grading the LOV. The ”hard-easy” effect, which means that subjective probability judgments are systematically changing from over- to underconfidence with decreasing task difficulty, is known [12]. Furthermore, variations between different investigators have to be assumed as well as an influence of personality characteristics on the estimation [13]. There are studies that show that accuracy of a diagnosis can be estimated, but overall physicians appear to be poorly calibrated, prone to over- and underestimation as well as poor in monitoring their performance [14–20]. This could limit potential benefits considerably. Therefore, it is also conceivable that the perception approach reveals just to be beneficial in small niches (e.g. When to ask for additional advice?). Stating difficulty and uncertainty can be interpreted as a sign of weakness or insufficient knowledge and/or skills and therefore the compliance of the decision-maker might be a critical topic.

A limitation of this communication concept is that an additional layer of explanation is necessary. For example, a diagnosis like dysplastic nevus, moderate atypia (D3) needs to be supplemented with information whether the investigator is considering a higher or lower grade of atypia as primary differential diagnosis. One might also standardize this point up to a certain extent but this will reduce the simplicity. For example, dysplastic nevus, moderate atypia (D3; max.) for maximum dysplastic nevus, moderate atypia (D3).

Synopsis

In the 18th century Voltaire wrote: "Doubt is not a pleasant condition, but certainty is an absurd one" [21]. It seems fitting that physicians are trained not to tolerate uncertainty, an unpleasant condition, especially with the burden of the responsibility for the lives and well-being of their patients. The issue is, however, of great importance considering the clinical reality with its substantial intra- and inter-observer variability in some fields. 2016 Simpkin et al. raised the question: "Tolerating Uncertainty — The Next Medical Revolution?" [22]. Here we provide a theoretical concept in which uncertainty, operationalized as LOV and Fusion Diagnoses, might lead to conscious purposeful acting. Of course, the suggested approach of quantifying uncertainty and including it into the diagnostic process is somewhat provocative. But this approach might have advantages because of its simplicity. In mesothelioma a classification according to the diagnostic uncertainty exists [23–25]. The concept is pragmatic and may suffer from predictable errors but it allows the communication of uncertainty in a defined way and with the conceivable potential to support physicians in their decisions. We therefore published this manuscript to encourage investigations that disadvantages and advantages can be elucidated.

For our opinion, the three main critical points of the concept are as follows: To what extent can the decision-maker really estimate the observer variability in the specific situations? Is the significance of likely disadvantages (for example, calibration and under- and overestimation) small enough that the advantages considerably overweight the disadvantages? Is it possible to convince physicians, with empirical evidence, to communicate their uncertainty in this way?

Conflict of interest

All authors declare that no conflict of interests exists.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mehy.2019.01.002.

References

[1] Bhise V, Rajan SS, Sittig DF, Morgan RO, Chaudhary P, Singh H. Defining and measuring diagnostic uncertainty in medicine: a systematic review. J Gen Intern Med 2017.

[2] Alam R, Cheraghi-Sohi S, Panagioti M, Esmail A, Campbell S, Panagopoulou E. Managing diagnostic uncertainty in primary care: a systematic critical review. BMC Fam Pract 2017;18(1):79.

[3] Attanoos RL, Bull AD, Douglas-Jones AG, Fligelstone LJ, Semararo D. Phraseology in pathology reports. A comparative study of interpretation among pathologists and surgeons. J Clin Pathol 1996;49(1):79–81.

[4] Galloway M, Taiyeb T. The interpretation of phrases used to describe uncertainty in pathology reports. Pathol Res Int 2011;2011:656079.

[5] Stoler MH, Schiffman M. Atypical squamous cells of undetermined significance-low-grade squamous intraepithelial lesion triage study (ALTS) group. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL triage study. J Am Med Assoc 2001;285(11):1500–5.

[6] Elmore JG, Barnhill RL, Elder DE, Longton GM, Pepe MS, Reisch LM, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. Br Med J 2017;357:j2813.

[7] Salerno SM, Alguire PC, Waxman HS. Competency in interpretation of 12-lead electrocardiograms: a summary and appraisal of published evidence. Ann Intern Med 2003;138(9):751.

[8] Martin B, Schäfer E, Jakubowicz E, Mayr P, Ihringer R, Anthuber M, et al. Level of interobserver variability estimation as a valuable tool: assessment of tumour budding in colon cancer. Histopathology 2018.

[9] Martin B, Schäfer E, Jakubowicz E, Mayr P, Ihringer R, Anthuber M, et al. Interobserver variability in the H&E-based assessment of tumor budding in pT3/4 colon cancer: does it affect the prognostic relevance? Virchows Arch Int J Pathol 2018;473(2):189–97.

[10] Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. Br Med J 1996;312(7023):71–2.

[11] Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980;302(20):1109–17.

[12] Suantak L, Bolger F, Ferrell WR. The hard-easy effect in subjective probability calibration. Organ Behav Hum Decis Process 1996;67(2):201–21.

[13] Schneider A, Wübken M, Linde K, Bühner M. Communicating and dealing with uncertainty in general practice: the association with neuroticism. PLoS One 2014;9(7):e102780.

[14] Uy RC, Sarmiento RF, Gavino A, Fontelo P. Confidence and information access in clinical decision-making: an examination of the cognitive processes that affect the information-seeking behavior of physicians. AMIA Annu Symp Proc AMIA Symp. 2014;2014:1134–40.

[15] Friedman CP, Gatti GG, Franz TM, Murphy GC, Wolf FM, Heckerling PS, et al. Do physicians know when their diagnoses are correct? J Gen Intern Med 2005;20(4):334–9.

[16] Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. Acad Med J Assoc Am Med Coll 2001;76(10 Suppl):S87–9.

[17] Meyer AND, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA Intern Med 2013;173(21):1952–8.

[18] Podbregar M, Voga G, Krivec B, Skale R, Pareznik R, Gabrscek L. Should we confirm our clinical diagnostic certainty by autopsies? Intensive Care Med 2001;27(11):1750–5.

[19] de Bruin ABH, Dunlosky J, Cavalcanti RB. Monitoring and regulation of learning in medical education: the need for predictive cues. Med Educ 2017;51(6):575–84.

[20] Davis DA, Mazmanian PE, Fordis M, Van Harrison R, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. J Am Med Assoc 2006;296(9):1094–102.

[21] Letter to Frederick William. Prince of Prussia, 1770 Nov 28. In: Tallentyre SG, editor. Voltaire in his letters. New York, NY, USA: G.P. Putnam's Sons; 1919. p. 232.

[22] Simpkin AL, Schwartzstein RM. Tolerating uncertainty – the next medical revolution? N Engl J Med 2016;375(18):1713–5.

[23] Ascoli V. Pathologic diagnosis of malignant mesothelioma: chronological prospect and advent of recommendations and guidelines. Ann Ist Super Sanita 2015;51(1):52–9.

[24] van Gelder T, Hoogsteden HC, Vandenbroucke JP, van der Kwast TH, Planteydt HT. The influence of the diagnostic technique on the histopathological diagnosis in malignant mesothelioma. Virchows Arch A Pathol Anat Histopathol 1991;418(4):315–7.

[25] Wertungsschema des europäischen Mesotheliom Panels [Internet]. [zitiert 29. September 2018]. Verfügbar unter: https://www.ruhr-uni-bochum.de/pathologie/mesotheliomregister/panel.html.