**ORIGINAL ARTICLE**

**A pilot study in ophthalmology of inter-rater reliability in classifying diagnostic errors: an underinvestigated area of medical error**

C E Margo

**Background:** Misdiagnosis is the least studied form of medical error. Before effective strategies to reduce misdiagnosis can be developed, there needs to be a better understanding of the factors that lead to these errors.

**Aim:** To evaluate the applicability and reliability of three classification systems for misdiagnosis.

**Design:** Retrospective independent analysis of five cases by clinical experts.

**Participants:** Three ophthalmologists trained in ocular oncology who devote at least 75% of their practice to ocular oncology.

**Main outcome measures:** Percentage agreement in determining cause of misdiagnosis.

**Results:** Participants agreed a misdiagnosis occurred in all cases and the error was graded as serious 1.4 of 15 times (93%). Inter-rater agreement for root cause varied among the three classification systems from 47% to zero.

**Conclusions:** Although there was excellent agreement among clinical experts of what constitutes serious misdiagnosis under idealized conditions, there is not a reliable method for categorizing the primary or root cause for these errors. The origins of misdiagnosis are complex, often multifactorial, and more difficult to categorize than other types of medical error. Misdiagnosis is a professional and public healthcare challenge that will require novel strategies to enable it to be successfully studied.

The debate over why medical errors occur has led to both greater awareness of the problem and renewed interest in how to prevent them. Despite the controversy over the accuracy of error estimates, several valid observations have emerged from this dialogue: (1) there is a need for better methods of error recognition and prevention; and (2) the problem of human error in medicine has not been effectively addressed in the past. While the reasons for these deficiencies are not entirely clear, a portion of the blame has been directed at cultural barriers in medicine that discourage discussion of error. Contributing to this controversy is the public perception that serious mistakes are tolerated by the medical profession.

Misdiagnosis is an important subset of medical error that has been largely ignored in the debate over quality improvement and error prevention. There is evidence to suggest that misdiagnosis is a widespread and under reported problem. Correlative autopsy studies have shown a consistently high frequency of misdiagnoses which have not declined with the introduction and dissemination of new medical technologies over five decades. A high proportion of misdiagnoses can be found from studies dealing with other subjects such as the economic impact of second medical opinions. Investigation into epidemics of misdiagnosis due to multiple time clustered errors of diagnosis reveals institutional and professional pressure not to report such events. As a consequence, it is likely that serious diagnostic errors are underestimated. Physicians who are tested using “real life” clinical simulations commit a substantial number of errors in clinical reasoning (resulting in misdiagnosis) at all stages of training and experience, suggesting that similar mistakes occur in clinical practice. Errors in diagnosis are commonly detected when physicians are tested using standardized patients with known disease. The differences in accuracy rates and reliability of diagnostic procedures reported from clinical trials and from general practice indicate that improvements can be achieved in the application of diagnostic methods in general practice. Finally, cognitive errors in diagnosis contribute to one fifth of lawsuits against physicians.

Although the study of human error in medicine is in its infancy, it is clear that the greatest challenge will be in understanding and preventing mistakes for highly cognitive processes like clinical reasoning. Any corrective approach to the reduction of human error would benefit from knowledge of primary or root causes. The purpose of this study was to test the application and inter-rater reliability of different systems to classify the primary cause of misdiagnosis. A valid classification system would facilitate communication among public health officials, medical educators, and quality improvement investigators, and could provide the foundation for effective corrective action.

**METHODS OF CLASSIFYING ERRORS OF DIAGNOSIS**

A review of the literature failed to uncover any systematic means of classifying the cause of diagnostic errors. Errors in diagnosis have traditionally been graded as major errors or class I errors, depending on whether they are a principal or underlying cause of death or whether they would have affected prognosis if detected during life, respectively. These grading systems give no insight into the cognitive or technical processes that lead to errors. The literature review revealed several systems that could be adapted with little or no modification for the purpose of causal classification.

The first classification system is used without modification from the Adverse Medical Event System for transfusion medicine. This method of error analysis is a prototype national event reporting system for blood centers and hospital transfusion services and is thought to have widespread error analysis application in industry and in...
Inter-rater reliability in classifying diagnostic errors

The development and application of this system for transfusion medicine has been previously described.\(^1\) Briefly, the classification of root causes of medical error are divided into latent errors and active errors (table 1). Latent errors are subdivided into technical and organizational. Active errors are categorized as external and behavioral, which is further subdivided into knowledge based, skill based, rule based, and other. While this system has enjoyed success in identifying root cause of error in the field of blood banking, it has not been tested for its applicability to errors in diagnosing disease.

The second classification system is based on the “dimensions” of professional competence as outlined by Epstein and Hundert (table 2). This classification system is an inventory of professional skills and behavior considered essential for the competent practice of medicine.\(^2\) Given that this inventory of cognitive, technical, and higher level integrative skills characterized the dimensions of competent medical practice, it could also be used to identify the fundamental deficiencies in physicians whose performance is deemed suboptimal. The dimension of professional competence as proposed by Epstein and Hundert was modified for use in this study by removing three categories (referred to as context, relationship, and affective/moral) because of the difficulty of measuring these behavioral characteristics. The 22 subdivisions in the remaining four categories (cognitive skills, technical skills, integrative skills, and habits of mind) were condensed into 16 to reduce conceptual overlap.

### Table 1: Adverse Medical Event System

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent</td>
<td>Errors that result from underlying system failure</td>
</tr>
<tr>
<td>Technical</td>
<td>Refers to equipment, physical installation, material, labels, forms, etc</td>
</tr>
<tr>
<td>External</td>
<td>Technical failures beyond control of investigating organization</td>
</tr>
<tr>
<td>Design</td>
<td>Inadequate design of equipment and related support material</td>
</tr>
<tr>
<td>Construction</td>
<td>Correct designs were not constructed properly</td>
</tr>
<tr>
<td>Materials</td>
<td>Material defects</td>
</tr>
<tr>
<td>Organizational</td>
<td>Failures beyond control of investigating organization</td>
</tr>
<tr>
<td>External</td>
<td>Human failures beyond control of organization</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Inability to apply existing knowledge to novel situation</td>
</tr>
<tr>
<td>Rule based</td>
<td>Qualifications</td>
</tr>
<tr>
<td>External</td>
<td>Individual not qualified by education or training for task</td>
</tr>
<tr>
<td>Coordination</td>
<td>Lack of task coordination within healthcare team</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Failure of transfer of knowledge or information to staff</td>
</tr>
<tr>
<td>Management</td>
<td>Internal decision to relegate safety decision to an inferior position</td>
</tr>
<tr>
<td>Culture</td>
<td>Collective failure and its attendant modes to function safely</td>
</tr>
<tr>
<td>Active</td>
<td>Errors or failures that result from human behavior</td>
</tr>
<tr>
<td>External</td>
<td>Human failures beyond control of organization</td>
</tr>
<tr>
<td>Knowledge</td>
<td>Inability to apply existing knowledge to novel situation</td>
</tr>
</tbody>
</table>

### Table 2: Professional Skills and Behavior Inventory*

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faulty cognitive skills</td>
<td>inadequate core knowledge, inadequate basic communication skills, ineffective information management, inability to apply knowledge to real world, inadequate use of tacit knowledge, inadequate use of personal experience, faulty abstract problem solving, unable to recognize gaps in knowledge, unable to generate appropriate questions, ineffective resource utilization</td>
</tr>
<tr>
<td>Faulty technical skills</td>
<td>inadequate physical examination, inadequate interpretation of special studies</td>
</tr>
<tr>
<td>Faulty integrative skills</td>
<td>Failure of habits of mind: inattention to detail, lack of critical curiosity, unwillingness to acknowledge or to correct errors</td>
</tr>
</tbody>
</table>

*Adapted from Epstein and Hundert.\(^2\)

The third system is borrowed from a paper by Kassirer and Kopelman\(^3\) who compiled a comprehensive list of faulty cognitive practices that lead to error in diagnosis (table 3). The purpose of their compilation was to identify and classify cognitive errors so that a deeper understanding of the epidemiology and causes of diagnostic error could be obtained.\(^3\) This so-called catalogue of faulty cognition was itself derived from the scientific literature and is based on the assumption that clinical problem solving and the cognitive process of scientific discovery are inherently similar.\(^3\) The five categories of cognitive diagnostic errors along with the definitions proposed by Kassirer and Kopelman were used without modification. Although the origin of these three classification systems differs, each appears to have potential value in studying the cause of diagnostic errors.

### Table 3: Catalogue of Cognitive Errors*

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faulty “triggering”</td>
<td>Faulty “trigger” or hypothesis generation by patient’s complaint or by initial findings</td>
</tr>
<tr>
<td>Faulty context formation</td>
<td>Inability to correctly frame clinical problem, includes faulty interpretation of test results, faulty causal model, over reliance on a clinical axiom, etc</td>
</tr>
<tr>
<td>Faulty information gathering</td>
<td>Failure to exclude other diagnostic possibilities, the final step in confirming the accuracy of a diagnosis</td>
</tr>
<tr>
<td>No fault error</td>
<td>Beyond control of physician</td>
</tr>
</tbody>
</table>

*Modified from Kassirer and Kopelman.\(^3\)
Box 1 Clinical summaries of five patients studied

Case 1
A young boy with an enlarging right lower eyelid mass underwent three biopsies over an 8 month interval. The biopsies each showed chronic inflammation with eosinophils and necrosis. The anatomical diagnosis was descriptive and included the comment “consistent with eosinophilic granuloma”. Treatment with external beam radiation on three separate occasions and several courses of a corticosteroid was unsuccessful. Usually eosinophilic granulomas are very sensitive to radiotherapy. When the pathology slides were reviewed elsewhere, the diagnosis of fungal cellulitis was made. The boy eventually lost his right eye, eyelids, facial skin and orbit.

Areas of concern: A boy lost nearly a third of his face because three biopsies were misinterpreted as eosinophilic granuloma and cultures of the inflamed tissue were never taken.

Case 2
A large pigmented subretinal mass was found between the equator and ciliary body in an elderly man several weeks after uncomplicated cataract surgery for an early lens opacity (cataract). A second opinion on the nature of the lesions was requested. Ultrasound examination of the eye lesion (standardized A scan and B scan) were obtained. The retina examination before surgery was described as normal. Following cataract surgery the ocular media were clear and the patient’s vision was excellent. The eye was enucleated several weeks later because of the suspicion of malignant melanoma. Pathological examination showed a localized hemorrhage beneath the retina (suprachoroidal hemorrhage). This type of hemorrhage would typically resolve within weeks without adverse sequelae.

Areas of concern: A man had a normal vision eye removed because fundus examination and ultrasound studies of a localized hemorrhage were misinterpreted as a melanoma. The history of a normal retina examination just weeks before did not alter the decision for surgery.

Case 3
A middle aged man presented with a slowly enlarging “spot” in the temporal field of vision of his right eye for 8 weeks. His visual acuity in the eye was good. The patient told that his eye examination was normal except for an age related cataract which was the cause of his visual disturbance. The patient sought a second opinion 3 weeks later. A large pigmented tumor suspicious for malignant melanoma was found in the back of the eye. Its location would explain the blind spot. The patient decided to have the right eye removed several weeks later. The globe showed moderately large malignant melanoma, 15 mm in diameter and 12 mm in height.

Areas of concern: An easily visible melanoma was not seen on dilated eye examination and the complaint of peripheral blind spot was incorrectly attributed to a minimal cataract.

Case 4
A healthy 2 year old boy was referred for evaluation of a dark pupil. The diagnosis of hyphema (blood inside anterior chamber of the eye) secondary to juvenile xanthogranuloma (benign tumor) of the iris was considered likely. A dermatologist found no skin lesions. An ocular ultrasound to examine the posterior part of the eye was interpreted as normal. The child underwent two surgical procedures to irrigate the blood from the eye and to treat secondary glaucoma (raised intraocular pressure). The possibility of retinoblastoma (malignant tumor of the retina) was considered after the second surgery. The eye was enucleated 5 months after presentation and showed malignant tumor cells at the surgical edge of the optic nerve and in the previous surgical procedure site.

Areas of concern: The spontaneous accumulation of blood in the eye of a young child was attributed incorrectly to a rare benign intraocular tumor. An ultrasound study of the eye was interpreted as normal when in fact the eye was filled with a malignant tumor.

Case 5
A middle aged woman in good general health presented with pain around her right eye which had been present for 2 months. Her eye examination revealed a bulging right eye (proptosis) and redness of the ipsilateral upper eyelid. A CT scan of the orbit and laboratory studies were obtained. She was treated for inflammation (so called inflammatory pseudotumor) with oral prednisone and, later, narcotic analgesics. A biopsy of the lacrimal gland 16 months after presentation showed an adenocarcinoma with spread around the nerves.

Areas of concern: A CT scan of a malignant tumor of the lacrimal gland and orbit was misinterpreted as inflammation. When orbital pain worsened, other diagnostic possibilities were not entertained.

involves investigation into all aspects of care potentially related to the error. Because much of this material is not contained within the medical record, participants were provided access to the technical, organizational, and environmental details of each case through an interview. The participants were able to request any potentially relevant information including—but not limited to—the training and professional experience of physicians, qualifications of technical support staff, quality and condition of medical instrumentation, and type of clinical practice (private practice group, private practice solo, academic medical center, general or referral practice, etc).

The definition of misdiagnosis used by the participants in the study was a “conclusion about an abnormal state of health that leads to a management decision resulting in an otherwise avoidable injury”. The distinction between degrees of error (minor, moderate, and serious) is subjective and was left to the discretion of the reviewer.

Agreement among participants for each system was calculated by dividing the number of agreed responses by the total number of responses; proportions were expressed as percentages. A high level of agreement between participants for a given system would indicate that errors of diagnosis can be reliably classified according to primary or root cause.

The study was conducted between February and July 2002.

RESULTS
The three participants agreed that a misdiagnosis had occurred in all five cases (100% agreement). The errors were judged as serious in 14 of 15 occasions (93%) and once as moderate (7%). One participant graded the error in case 3 as
Table 4  Agreement in classification of primary errors of diagnosis

<table>
<thead>
<tr>
<th>Case</th>
<th>Reviewer</th>
<th>Adverse Medical Event System</th>
<th>Professional Skills and Behavior Inventory</th>
<th>Catalogue of Cognitive Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Beyond control</td>
<td>Inadequate interpretation*</td>
<td>Faulty verification</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Failure to monitor</td>
<td>Inadequate interpretation*</td>
<td>Faulty information gathering</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Incomplete assessment</td>
<td>Inadequate examination</td>
<td>Faulty hypothesis</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Incomplete assessment</td>
<td>Inadequate interpretation</td>
<td>Faulty interpretation</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Failure to monitor</td>
<td>Unable to recognize gaps</td>
<td>Failure to exclude diagnosis</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>Incomplete assessment</td>
<td>Inadequate examination</td>
<td>Faulty hypothesis</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>Failure beyond control</td>
<td>Inadequate examination</td>
<td>Faulty hypothesis</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Incomplete assessment</td>
<td>Inadequate examination*</td>
<td>Faulty interpretation</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Decision to relegate</td>
<td>Inadequate use personnel</td>
<td>Failure to exclude diagnosis</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>Incomplete assessment</td>
<td>Inadequate interpretation</td>
<td>Failure to exclude diagnosis</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Incomplete assessment</td>
<td>Inadequate interpretation*</td>
<td>Inadequate examination*</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>Incomplete assessment</td>
<td>Inadequate interpretation</td>
<td>Faulty hypothesis</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>7/15 agreements (46.6%)</td>
<td>5/15 agreements (33.3%)</td>
<td>No agreements (0%)</td>
</tr>
</tbody>
</table>

*Case agreement.

DISCUSSION

There was near uniformity of agreement (93%) that the errors of diagnosis in these five cases were serious. These cases are not rare events. Current analysis of autopsy data indicates that a major diagnosis goes undetected in more than 74,000 individuals who die in hospitals in the United States each year. Nearly 35,000 of these patients could have survived hospital discharge if the error had not occurred.

Why do diagnostic errors occur and can such errors be prevented in the future?

There have been few formal investigations or discussions into the causes of misdiagnosis despite its importance to patient safety.

The three clinical experts in this study agreed on the primary cause of error, occasionally using three different systems for classification. Two of the systems in this study (Adverse Medical Event System and Professional Skills and Behavior Inventory) showed some potential usefulness but, even then, all three experts could agree only once with each system (table 4). Two of the three experts agreed in two other cases using the Medical Event Reporting System and in one case using the Professional Skills and Behavior Inventory. The diversity of opinion offered by these three experts may even be more noteworthy than their modest agreement.

Two of the systems in this study (Adverse Medical Event System and Professional Skills and Behavior Inventory) showed some potential usefulness but, even then, all three experts could agree only once with each system (table 4). Two of the three experts agreed in two other cases using the Medical Event Reporting System and in one case using the Professional Skills and Behavior Inventory. The diversity of opinion offered by these three experts may even be more noteworthy than their modest agreement.

The rather disappointing results of this study raise the question of why such a wide variety of primary causes were ascribed to each case by clinical experts. There is no simple explanation. Perhaps the categorical choices were too abstract or too conceptually overlapping. Maybe the understanding of medical decision making itself is too rudimentary. On the other hand, the multidimensional nature of decision making may not be easily categorized. Studies of complex work environments have shown that, once a minor mistake enters a system, secondary errors can propagate the problem and increase the likelihood of serious failure, particularly when no safeguards are built into the system to limit damage. Secondary errors—because of their immediate clinical consequences—can often eclipse primary mistakes in importance to the clinical outcome. The traditional domains used to categorize cause of injury such as human (behavioral), organizational, and technical also have considerable overlap in clinical practice which makes categorical distinctions potentially ambiguous.

There have been few analytical studies of diagnostic errors. Battles and Shea performed an analysis of the root cause of medical error in a teaching hospital using the Eindhoven Classification Model (similar to the Adverse Medical Event System employed in this study). They found that root cause analysis can be a valuable source of information for guiding educational and system change in the hospital. Their study, unlike this one, dealt with diverse types of medical error and did not focus on mistakes due to diagnosis.

The definition of misdiagnosis in this study is based on a medical injury model and not on a medical error model. Error and harm are not always linked. The medical error model, which includes mistakes that do not harm patients, would create too broad a definition for the purposes of studying misdiagnosis. The normal process of making a clinical diagnosis involves hypothesis testing. So called

Key messages

- Errors in diagnosis belong to a subset of medical error that has been largely ignored.
- There is compelling evidence from a variety of sources to indicate that errors of diagnosis are an important public health concern.
- Three classification systems to classify the cause of diagnostic errors were tested for inter-reviewer reliability.
- Although there is good agreement among experts about what constitutes a serious error in diagnosis, classification of primary or root cause is not reliable using available systems.
- Reduction of errors in diagnosis will improve medical care. Progress towards this goal will depend on a better understanding of the primary causes of these errors.
working or provisional diagnoses are often corrected as a clinical work up proceeds through logical stages. While there is considerable controversy over the roles served by the medical injury and medical error models in improving patient safety, the medical injury model has the practical advantage of establishing a less controversial threshold for the study of misdiagnosis by circumventing medical errors that cause no harm. Because of physician sensitivity and concerns over medicolegal liability, pilot studies of diagnostic error would be better received if based on the medical injury model.

The rationale for this study is based on the assumption that the opportunity to detect preventable morbidity and mortality increases when medical errors are defined in terms of failed processes rather than personal deficiencies. The results of this study indicate that, in an idealized situation (in which concerns over personal and professional reputation and medical liability are removed), experts have difficulty in agreeing upon the primary cause(s) of serious diagnostic error. Although this was a small study, it suggests that sentinel event tracking and the current injury prevention models that rely on root cause analysis might not be well suited for dealing with errors of diagnosis. The causal web leading to patient injury from errors in diagnosis may, in fact, be too complex to be deciphered using traditional methodologies.

While this pilot study failed to find a reliable means to classify diagnostic errors, it did confirm the ability of clinical experts to reliably identify serious diagnostic errors when they are made. Greater awareness of diagnostic errors may serve as a stimulus for others to pursue research in this area. Without a reproducible system to categorize diagnostic errors, there cannot be an effective means of studying the diversity and magnitude of the problem. The “art” of clinical diagnosis should no longer be considered beyond quality of care scrutiny. The medical profession needs to take an active role in studying the underlying causes of diagnostic error.

ACKNOWLEDGEMENTS

Drs D S Bartenstein, M W Wilson, and Z A Kargioglu generously gave their time to participate in this study.

REFERENCES

2. Leape LL. Institute of Medicine medical error figures are not exaggerated. JAMA 2000;284:95–7.

Case record review discussion—have your say!

• Much information about the incidence of adverse events is derived from case record review (CRR). In this issue Woloshynowych and colleagues (pp 411–415) recommend a format for improving CRR. But one of the limitations of CRR is the quality of case records. As Wilson points out (pp 402–403), improving the case records would further improve CRR. Some standards for case records exist, but even then it is not clear how well these are followed. What are your views about the quality of case records? How can they be improved?

Have your say. Enter the discussion about case records on the journal website.

• What are the key functions of the medical record?
• What are the key elements that would need to be included in a medical record in order to achieve these functions?
• How do we implement the recommendations from two above?
A pilot study in ophthalmology of inter-rater reliability in classifying diagnostic errors: an underinvestigated area of medical error

C E Margo

Qual Saf Health Care 2003 12: 416-420
doi: 10.1136/qhc.12.6.416

Updated information and services can be found at:
http://qualitysafety.bmj.com/content/12/6/416

These include:

References
This article cites 21 articles, 2 of which you can access for free at:
http://qualitysafety.bmj.com/content/12/6/416#BIBL

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/