Diagnostic categorization in EQA schemes

Nicholas P Mapstone

Abstract

The different ways in which diagnoses may be categorized in EQA schemes are considered. They may be grouped in very large diagnostic categories, or in narrow specific diagnoses. They may be categorized as dangerous or non-dangerous diagnoses. They may be categorized by the participants or a panel of experts. They may be categorized according to organ system, but rarely by underlying pathological process. The ways in which these categories are reached is considered, and the effect this has on the educational role of EQA schemes. This categorization will also influence the participant's strategy to ensure that they are not categorized as a poor performer.

Keywords diagnostic errors/classification; healthcare; pathology/standards; quality assurance; quality control

Introduction

External Quality Assurance (EQA) is now widely accepted by histopathologists in the UK.[1] They use it to demonstrate that their diagnostic powers are in line with the majority of their colleagues. It is likely that pathologists will use EQA as one way to fulfil the requirements of revalidation. Much of the process of EQA is in the categorization of cases. This short article will consider that categorization process, and how it affects EQA scoring.

Determination of the 'correct' diagnosis

All cases which are used in EQA will have a 'correct diagnosis' against which participants' diagnoses are compared. Most cases will be submitted by the original reporting pathologist with their original diagnosis. That is one person's opinion and may not be correct. So the 'correct diagnosis' will be determined either by an expert panel or by the most popular diagnosis amongst all participants.

Popular diagnosis schemes

Most EQA schemes have more than 50 participants. Obviously the more participants in a scheme, the more possibility for variety in diagnoses. Occasionally there will be only one diagnosis. Some would consider such a case to be too 'easy', but most schemes will have one or two cases per circulation with such excellent consensus. It is much more common to get multiple diagnoses and often there will be almost as many diagnoses as participants. Many of the differences in those diagnoses will be differences in terminology.

The first step for a scheme organizer, having collected all of the participants' diagnoses, is to group them into broad diagnostic categories to facilitate the choice by the participants of the correct diagnosis. The organizer has, at this stage, to decide which differences are merely terminological and which are of diagnostic importance. A classification error at this stage – grouping together diagnoses which later turn out to be significantly different, can cause massive duplication of work, having to rescore all participants' responses.

A computer program has been written by Professor Peter Furness[2] to facilitate this compilation process and allows a total of 10 diagnostic categories for each case. It gives a weighting to each of those 10 categories depending upon how many participants made the diagnosis, and how confident of it they were (Table 1).

The scheme organizer has to exercise some other judgements about the case. If two diagnoses for one case have been offered by many participants, how is that categorized? For example, an appendix might show inflammation and worms. Does the organizer group everyone by the mention of inflammation? Or should there be two groups? 'Inflammation and worms', versus 'Inflammation and no worms mentioned'. Some participants may have considered the worms, or spirochaetes or whatever, as being not worth mentioning. With two potential diagnoses in a case there are four possible diagnostic groups. With three potential diagnoses (faecolith, appendicitis, worms) the number of categories increases exponentially. Many participants will (not unreasonably) describe everything that they can see on a slide.

In this example, less than half of participants made the most popular diagnosis of sessile serrated polyp or sessile serrated lesion (diagnosis 1). There is no consensus as to the main diagnosis. A few participants wrote both hyperplastic polyp and sessile serrated polyp as their main diagnosis (diagnosis 4). Only one participant suggested serrated carcinoma as a possible diagnosis, and they only scored it as 1/10, giving 9/10 to some other diagnosis

	Circulation: R	**Case number: 804**	**Number of responses: 214**	**Date: 3 Nov 10**
	Diagnostic categories:		**Score**	
1	Sessile serrated polyp/lesion		4.41	
2	Serrated adenoma		1.70	
3	Sessile serrated adenoma		2.10	
4	Hyperplastic polyp/sessile serrated polyp		0.21	
5	Hyperplastic polyp		1.57	
6	Normal		0.01	
7	Serrated carcinoma		0.00	

Highest scoring diagnosis was 1 with 4.41 Asterisks (if any) indicate dangerous diagnoses.

Table 1

Nicholas P Mapstone *FRCPath is a Consultant in the Department of Histopathology at Royal Lancaster Infirmary, Lancaster, UK.* *Conflicts of interest: the author is chair of the 'NQAAP for Histopathology incorporating the Steering Committee for Interpretive EQA', and the organiser of a gastrointestinal pathology EQA scheme, and the Bowel Cancer Screening Program EQA scheme.*

The scheme organizer has to decide which diagnostic categories need to be considered by all the participants, and which can be safely ignored to simplify categorization. What is the diagnostic crux of the case?

This difficulty means that it is rare to require grading or staging diagnoses in EQA schemes. A case of gastrointestinal stromal tumour may result in a range of differential diagnoses, and prognosis for the GIST may complicate the diagnostic categories beyond what is practical.

Once the draft diagnostic categories have been compiled by the organizer, they are presented to a group of the participants — either at a meeting or in a discussion facilitated by, for example, email. Participants are presented with the (maximum of 10) diagnostic categories for each case and asked to decide which categories should be lumped together. There is no facility for splitting at this stage — unless the scheme organizer goes back to all the responses and reclassifies them. The participant in the back of the room who sees their diagnosis alone in a separate category now has a definite interest in getting it included with the most popular diagnosis — their score will depend upon how many of their colleagues made the same diagnosis. This is where some of the most heated and prolonged discussions can occur around EQA. If the most popular diagnosis is 'collagenous colitis', then participants who have diagnosed microscopic colitis will argue that their less specific diagnosis is included with the main diagnostic group — which will perhaps change its name to 'microscopic colitis, incorporating collagenous colitis'. The participant who has diagnosed 'inflammation' or 'colitis' might make the same argument — with less likelihood of success (Table 2).

Following this diagnostic lumping process, if fewer than 80% of participants have been grouped together in the main diagnostic category, the case is usually excluded as not having sufficient consensus.

Expert panel schemes

The process of identifying the most popular diagnosis and thrashing it out at a participants meeting is cumbersome. Identifying a panel of one or more experts who make the definitive diagnosis is much simpler but perhaps less popular with the body of participants. Problems of diagnostic categorization will be much reduced. Participants will usually receive a score of one if they get the case 'correct' and none if 'incorrect'.

Whilst there are arguments for both ways of identifying the 'correct' diagnosis, it is rare that participants as a group will come to an obviously incorrect consensus diagnosis. It is much commoner for such a case to result in a wide spread of diagnoses, and thus to be excluded from scoring due to the absence of consensus.

In the example seen in Table 1, following the participants meeting, those diagnoses which used the term serrated (except for carcinoma) have been combined. There is now consensus as to the main diagnosis

	Circulation: R	Case number: 804	Number of responses: 214	Date: 9 Nov 10
	Diagnostic categories		Score	
1	Sessile serrated lesion		8.42	
5	Hyperplastic polyp		1.57	
6	Normal		0.01	
7	Serrated carcinoma		0.00	

Highest scoring diagnosis was 1 with 8.42 Asterisks (if any) indicate dangerous diagnoses.

Table 2

Dangerous diagnoses

Sometimes participants will lose marks for diagnoses whose difference from the most popular diagnosis is minimal or even terminological. Conversely occasional incorrect diagnoses are so drastically incorrect that it seems unfair that the penalty is the same as for a terminological error. There are advocates for the separate category of 'dangerous diagnosis'. This category has fallen into disfavour at least in part because of the difficulty of defining a 'dangerous diagnosis'. Overdiagnosing carcinoma would likely be a dangerous diagnosis. Missing amoebae in colitis might also be considered dangerous. Missing giardiasis might have major consequences for a patient — but would probably not be called 'dangerous'.

Diagnostic granularity

Most schemes require participants to produce a classical histological diagnosis for each case. The participants, or a panel, then combine those diagnoses into diagnostic categories as we have seen. The breast[3] and bowel cancer screening programme EQA schemes however ask participants to classify cases into broad diagnostic categories — e.g. Non-neoplastic, low-grade dysplasia, high-grade dysplasia, malignant. This has many advantages, which are mainly administrative. It makes it much easier to compile large numbers of responses from hundreds of participants. It makes the process of identifying the correct diagnosis — usually by a panel of experts, much easier. It also allows a participant a semiquantative way of seeing how far away they might be from the 'correct' diagnosis.

This difference, between the non-granular diagnoses of the screening EQA schemes and the much more granular diagnoses in the other schemes, reflects a major philosophical difference in the rationale of EQA schemes. It is often said that the main role of EQA is educational, although many consider its quality assurance role as predominant.

The educational function is best served by having classical, specific diagnoses. This allows diagnostic variants to be included and assessed. Missing a rare diagnosis in an EQA scheme means it is less likely to be missed in routine diagnosis.

The assessment function is best served by having broad diagnostic categories. The spread of diagnoses will necessarily be smaller, but when aberrant diagnoses occur they are more obvious and objective. These schemes still have an educational role, but it is less pronounced. In the bowel cancer screening programme scheme, participants have a single box to tick if they think a specimen is not dysplastic — they do not have to identify Peutz Jeghers, hyperplastic or juvenile polyps. From the bowel cancer screening programme's point of view that is not a crucial distinction.