Assessing the Generalizability of Prognostic Information

Amy C. Justice, MD, PhD; Kenneth E. Covinsky, MD, MPH; and Jesse A. Berlin, ScD

Physicians are often asked to make prognostic assessments but often worry that their assessments will prove inaccurate. Prognostic systems were developed to enhance the accuracy of such assessments. This paper describes an approach for evaluating prognostic systems based on the accuracy (calibration and discrimination) and generalizability (reproducibility and transportability) of the system's predictions. Reproducibility is the ability to produce accurate predictions among patients not included in the development of the system but from the same population. Transportability is the ability to produce accurate predictions among patients drawn from a different but plausibly related population. On the basis of the observation that the generalizability of a prognostic system is commonly limited to a single historical period, geographic location, methodologic approach, disease spectrum, or follow-up interval, we describe a working hierarchy of the cumulative generalizability of prognostic systems.

This approach is illustrated in a structured review of the Dukes and Jass staging systems for colon and rectal cancer and applied to a young man with colon cancer. Because it treats the development of the system as a "black box" and evaluates only the performance of the predictions, the approach can be applied to any system that generates predicted probabilities. Although the Dukes and Jass staging systems are discrete, the approach can also be applied to systems that generate continuous predictions and, with some modification, to systems that predict over multiple time periods. Like any scientific hypothesis, the generalizability of a prognostic system is established by being tested and being found accurate across increasingly diverse settings. The more numerous and diverse the settings in which the system is tested and found accurate, the more likely it will generalize to an untested setting.

This paper is also available at http://www.acponline.org.

Ann Intern Med. 1999;130:515-524.

From Pittsburgh Veterans Affairs Health Care System and University of Pittsburgh, Pittsburgh, Pennsylvania; University of California, San Francisco, and San Francisco Veterans Affairs Medical Center, San Francisco, California; and School of Medicine and Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, Pennsylvania. For current author addresses, see end of text.

Your secretary just called to say that a favorite patient of yours (a 45-year-old high school teacher) with colon cancer dropped off his surgical and pathologic reports. She reminds you that he is scheduled this afternoon for a second opinion about his prognosis. Scanning the reports, you note that the surgeon staged the cancer at Dukes stage C1 on the basis of negative margins and 1 positive out of 28 lymph nodes and that the pathologist staged the microscopic tissue at Jass stage IV. You are not sure how to translate these stages into useful prognostic information for your patient.

Time is short. You log on to MEDLINE, type "Dukes and Jass," and select the years 1966 to 1997. The combined terms generate 18 references (1–18). Of these, 4 are independent reports of observed mortality rates (8, 11, 17, 19). You also track down the original reports of the Dukes and Jass systems (20, 21) and learn that both systems were developed at St. Mark's Hospital in London more than 30 years ago (20, 21). The Dukes system is based on histology and extent of local, lymphatic, and venous spread seen in the surgical specimen (20), and the Jass system is based on microscopic pathologic staging (21). However, the reported mortality rates by stage for these systems vary widely. How do you tell which report and which system are most likely to pertain to a 45-year-old teacher from Cleveland, Ohio?

Physicians are frequently asked for prognostic assessments and often worry that their assessments will prove inaccurate (22, 23). Prognostic systems, including risk factors, staging systems, decision rules, statistical models, and computer algorithms, have been developed to standardize and enhance the accuracy of prognostic assessments (24, 25). Although diverse techniques are used to develop these systems, all use a sample of patients for whom the outcome is known to relate baseline characteristics to an outcome of interest. Once a system is developed, it can be used to generate predictions for patients whose outcome is not yet known. A common problem in the application of prognostic systems is that the accuracy of the predictions degrades from the sample in which the system was first developed to subsequent application; that is, the systems do not generalize (26).

Although much has been written on the evalua-

Table 1. Definitions of Accuracy and Generalizability

Term	Definition or Criteria
Accuracy	The degree to which predicted outcomes match observed outcomes
Calibration	Predicted probability is neither too high nor too low (commonly shown with calibration curves)
Discrimination	Relative ranking of individual risk is in correct order (observed event rates in those with higher scores are higher); commonly measured with the area under the receiver-operating characteristic curve
Generalizability	Ability of a prognostic system to provide accurate predictions in a new sample of patients
Reproducibility	The system is accurate in patients who were not included in development but who are from an identical population
Transportability	The system is accurate in patients drawn from a different but related population or in data collected by using methods that differ from those used in development
Historical	Accuracy is maintained when the system tested in data from different calendar time
Geographic	Accuracy is maintained when the system is tested in data from different locations
Methodologic	Accuracy is maintained when the system is tested in data collected by using different methods
Spectrum	Accuracy is maintained in a patient sample that is, on average, more or less advanced in disease process or that has a somewhat different disease process or trajectory
Follow-up interval	Accuracy is maintained when the system is tested over a longer or shorter period

tion and reporting of prognostic systems (25, 27–40), few investigations have directly addressed the issue of generalizability, also known as *external validity* (29). Instead, discussion has focused on evaluating issues of internal validity, such as the sample in which the system was developed, the variables used in the model, the techniques used in system development, or the accuracy of the system in the sample in which it was developed. These factors offer important insights into the probable generalizability of the system to a new sample of patients, but they do not directly test subsequent performance (25).

We discuss the importance of systematically testing the subsequent performance of a system. We begin by defining the relation between accuracy and generalizability, components of accuracy (calibration and discrimination), and components of generalizability (reproducibility and transportability). We then discuss issues of transportability and propose a five-level hierarchy of external validity based on the type and degree of transportability tested. We illustrate this approach with a structured review of the Dukes and Jass staging systems for colon and rectal cancer as applied to the 45-year-old teacher described previously. Because this approach treats prognostic system development as a "black box" and focuses on subsequent performance, it can be applied to any prognostic system, no matter how complex.

Accuracy and Generalizability

Accuracy and generalizability are related concepts (**Table 1**). *Accuracy* is the degree to which predictions match outcomes. *Generalizability* is the ability of the system to provide accurate predictions in a different sample of patients.

Components of Accuracy

A series of numeric predictions may be inaccurate in two ways. The predicted probability may be too high or too low (an error in *calibration*), or the relative ranking of individual risk may be out of order (an error in *discrimination*). Assume that we have observed a sample of patients with colon cancer for whom the overall 5-year mortality rate was 50%. A system that predicted a 50% probability of death at 5 years for each patient would be perfectly calibrated. However, it would not discriminate among patients who lived and those who died within the interval. Conversely, a system that assigned a 10% probability of death at 5 years to patients who lived and an 11% probability to those who died would be perfectly discriminating but poorly calibrated.

The relative importance of calibration and discrimination depends on the intended application. If predictions are used to counsel a patient, the accuracy of the numeric probability (calibration) is important. Patients are not concerned about how sick they are relative to other patients with the disease; instead, they are concerned with the likelihood that their disease will result in death or some other important outcome (such as, in the case of our hypothetical patient, the inability to handle the challenges of teaching) within a defined period of time. Calibration is also important in health services research. When predicted and observed mortality rates are compared to identify unexpectedly high or low rates, errors in calibration can cause large numbers of hospitals or providers to appear to have excessively high or low rates of mortality when, in fact, the model is not calibrated (41). In contrast, if predictions are used to stratify patients by stage of severity in order to compare treatments within a given stage, the important aspect of accuracy is whether patients whose disease is within a stage are equally likely to experience the outcome and that the stages are correctly ranked in order of risk (discrimination) (41).

Calibration and discrimination are evaluated in different ways. Calibration is not routinely measured but can be illustrated by using calibration curves, which plot predicted versus observed outcomes (42). Discrimination is commonly measured by using the area under the receiver-operating characteristic (ROC) curve (43). This area ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination) and reflects the probability that in all possible pairs of patients in which one patient lives and one dies, a higher risk is assigned to the patient who died than to the one who lived. The area under the ROC curve can be directly calculated from a table of observed and predicted outcomes (44). It can also be calculated for continuous prognostic estimates (with or without censored observations) by using the C statistic (37). Several sources provide more thorough discussion of measures of calibration and discrimination (37, 41–43, 45–49).

Components of Generalizability

No matter how calibrated and discriminating a system may be in development, a system that can only predict outcomes in the sample in which it was developed is useless (25, 50). For a system to be generalizable, the accuracy (that is, calibration and discrimination) of the system must be both reproducible and transportable.

Reproducibility

Reproducibility requires the system to replicate its accuracy in patients who were not included in development of the system but who are from the same underlying population. A test of the reproducibility of the system evaluates the degree to which the system is fit to real patterns in the data rather than to random noise. The system is more likely to be fit to random noise (*overfit*) when the ratio of the number of variables to the number of patients experiencing events is small (37). If a prognostic system is overfit, it may not generalize well.

Methods for evaluating the reproducibility of a prognostic system have been thoroughly described elsewhere (32, 35, 51, 52) and are based on the use of data resampling techniques (such as bootstrapping) to evaluate the degree of overfitting. Bootstrapping techniques can evaluate errors in discrimination and in calibration and are particularly important when the sample used to develop the model is small (35, 52).

Transportability

Alternatively, a system may be reproducible (that is, perform well when tested by bootstrapping) and degrade in subsequent patient samples because of *underfitting* (37, 53). Underfitting occurs when important independent predictors of outcome are omitted from the system. For example, a system for breast cancer that omits the presence of metastasis may perform well in a sample of patients with no metastatic disease and degrade badly when it is tested in a more diverse sample. Bootstrapping of the development sample would not detect this problem because the development sample is homogeneous with respect to metastatic disease. Omission of metastasis in breast cancer staging is, of course, an obvious mistake; however, not all important prognostic variables in a disease state are known. Because any given sample may be homogeneous for an important variable (known or unknown), it is important to test the system in subsequent samples.

Transportability requires the system to produce accurate predictions in a sample drawn from a different but plausibly related population or in data collected by using slightly different methods than those used in the development sample. In the case of our patient, we want to know whether staging systems developed at St. Mark's Hospital in London more than 30 years ago pertain to a 45-year-old teacher from Cleveland whose disease was staged at Case Western Reserve University Hospital in 1998. Systems that are underfit may demonstrate reproducibility but not consistent transportability (50). Therefore, data from a separate, nonidentical sample are needed to assess transportability (50).

The specific ways in which the new sample or data collection technique differs from that used in the development of the system determine the components of transportability to be tested. We discuss five types of transportability: historical, geographic, methodologic, spectrum, and interval. In many of the following examples, we used standard methods (44) to calculate the area under the ROC curve on the basis of data from published reports. Transportability can have degrees: A system that passes a mild test may fail a more extreme test. The type and degree of transportability that are most important depend on the intended application of the system.

Historical transportability requires that the system maintain accuracy when it is tested in cohorts from different historical periods. It is particularly important when the severity of the disease is likely to have changed over time. This occurs when patients' conditions are diagnosed earlier owing to more aggressive screening practices or when medical treatments become more effective. For example, a system previously developed by Justice and colleagues (54) was good at discriminating outcomes in an early cohort of hospitalized patients with AIDS from 1981 to 1987 (area under the ROC curve, 0.81 [95% CI, 0.73 to 0.89]). However, its discrimination degraded substantially in two subsequent cohorts from 1987 to 1988 and 1990 to 1991 (areas under the ROC curve, 0.68 [CI, 0.61 to 0.75] and 0.65 [CI, 0.60 to 0.70], respectively) (55, 56). You, in turn, might ask whether staging systems developed more than 30 years ago pertain to a patient whom you will see this afternoon.

Geographic transportability requires that the sys-

Table 2. **A Hierarchy of External Validity for Predictive Systems**

Level of Validation	Cumulative Generalizability Evaluated
0. Internal validation	Reproducibility
1. Prospective validation	Reproducibility, historic transportability
2. Independent validation	Reproducibility, historic transportability, geographic transportability, methodologic transportability, spectrum transportability
3. Multisite validation	Reproducibility, historic transportability, geographic transportability, methodologic transportability, spectrum transportability
4. Multiple independent validations	Reproducibility, historic transportability, geographic transportability, methodologic transportability, spectrum transportability
5. Multiple independent validations with life-table analyses	Reproducibility, historic transportability, geographic transportability, methodologic transportability, spectrum transportability, follow-up period transportability

tem remain accurate when it is tested in data from other locations. Mocroft and colleagues (57) developed a system to predict outcomes in patients with AIDS that discriminated reasonably well among three patient populations that were treated at hospitals in London (areas under the ROC curve 0.88 [CI, 0.83 to 0.92], 0.93 [CI, 0.67 to 1.0], and 0.89 [CI, 0.86 to 0.95]) (57). However, the system was substantially less discriminating when it was tested in an Italian cohort (area under the ROC curve, 0.71 [CI, 0.56 to 0.76]) (58). Surprisingly, the authors of the Italian study thought that their data validated the usefulness of the Mocroft system rather than questioning its generalizability. You, in turn, might ask whether systems developed in London will work in a patient from Cleveland.

Methodologic transportability requires that the system maintain accuracy when it is tested in data collected by using alternative methods. Charlson and colleagues (36) have speculated that this may be the most common problem when systems fail to generalize. Differences in the methods used to define variables and collect data can cause substantial differences in the performance of the system. If the system cannot be reproducibly applied by other investigators, we must question its generalizability. If, however, these differences are in clear violation of the recommended method, the validation may not have been a fair test of the system. An example of the latter case is an extreme test of the methodologic transportability of the Charlson Comorbidity Index. Although this index was developed by using medical record data with good discrimination in the development and initial validation sample (areas under the ROC curve, 0.82 [CI, 0.78 to 0.87] and 0.98 [CI, 0.95 to 1.0], respectively) (26), it was tested by using discharge diagnosis codes; this resulted in an alarming decay in discrimination (area under the ROC curve, 0.57 [CI, 0.46 to 0.64]) (59). You might ask whether systems based on surgical specimen and pathologic analyses completed at St. Mark's Hospital in London generalize to methods practiced by surgeons and pathologists who practice in Cleveland, Ohio.

Spectrum transportability refers to the ability of the system to generate calibrated and discriminating predictions in patients who are, on average, more (or less) advanced in their disease process or who have a somewhat different disease. The importance of assessing the accuracy of predictive systems in patient samples of varying disease spectrum has been carefully examined with regard to diagnostic systems (60–62), and two concerns apply to prognostic systems as well. First, a different average level of severity suggests a different overall rate of outcome, which may substantially alter the calibration of the system. A perfect system should be able to sort patients according to risk for death so that calibration would not degrade in samples of diverse outcome prevalence. In reality, the calibration of most systems will be compromised when these systems are tested in a sample of patients with very different levels of disease severity (41). This is because statistical models are calibrated to the overall outcome prevalence in the development sample. Discrimination may also be altered. Perhaps the most difficult test of discrimination occurs when the spectrum of disease narrows from both sides; that is, the test sample includes many patients who have an illness of intermediate severity and very few who are either severely ill or not very ill at all.

Follow-up period transportability requires that the system maintain accuracy when predictions are tested over a longer or shorter follow-up period. Follow-up period transportability and spectrum transportability are closely related. Alterations in both disease spectrum and follow-up period can change the prevalence of the outcome and may thereby alter the calibration of the system. Furthermore, discrimination may degrade when the system is tested for follow-up period transportability. If the model fails to identify variables that are important both to short- and long-term outcomes, performance may degrade when the model is tested over different follow-up periods. For example, your patient may want to know about survival periods other than 5 years. Can these systems also predict for shorter or longer periods?

A Hierarchy of Prognostic Validation

We have identified types of generalizability and a method by which each type of generalizability can be tested (by measuring the accuracy of the system in a different sample of patients). We propose a five-level hierarchy of external validation for prognostic systems (**Table 2**). The level of external validation reflects the cumulative types of generalizability that have been tested and the degree of accuracy (calibration and discrimination) maintained in these cumulative tests. A system can never be fully validated—you can never be certain that it will apply to the next patient to come into your office. Nevertheless, like any other scientific hypothesis, the external validity of a prognostic system is established by being tested and found accurate across increasingly diverse settings. The more diverse the previous settings in which the system has been tested and found to be accurate, the more likely it is that the system will generalize to an untested setting. Thus, this hierarchy views external validation as an iterative, cumulative process. The hierarchy we provide is not exhaustive; it is intended instead to address commonly used study designs.

0. *Internal validation* tests the accuracy of the system in the sample used to develop the system. It is commonly restricted to data from a single geographic site. Internal validation uses data exclusion techniques (random split sample) or resampling techniques (bootstrapping) (32, 35, 36, 51, 52) to determine the degree to which the model may be overfit. Thus, it evaluates reproducibility alone and provides no direct test of external validity. Nevertheless, internal validity is often a prerequisite for external validity (63).

1. *Prospective validation* tests the accuracy of the system in data collected after development of the system. It is typically carried out by the same investigators at the same institution. It evaluates reproducibility and tests the susceptibility of the system to mild differences in historical time frame (historical transportability). However, it rarely tests methodologic or substantial geographic transportability.

2. *Independent validation* tests the accuracy of the system in data collected by independent investigators, usually at a different site. The study sample is almost always drawn from a different historical time frame. Establishing that the system is equally accurate when applied by independent investigators is important because independent investigators are unlikely to study identically selected patients and are unlikely to have similar idiosyncrasies in data collecting and recording techniques (36).

3. *Multisite validation* tests the accuracy of the system at multiple geographic sites. It directly tests geographic transportability and may also provide a mild test of methodologic transportability. Just as in other research domains, many investigators develop and test their prognostic systems by using data from multiple sites (64–67). However, investigators rarely report the variation in results by geographic locale. We encourage investigators to report variation in their results by geographic site (where sample size is adequate to avoid extremes in random variation

Table 3. Discrimination in Development and Validation Studies: The Dukes and Jass Systems

System	Investigator (Reference)	Site	Years	Disease Spectrum	Overall Mortality Rate, %	Area under the Receiver-Operating Characteristic Curve*		
						5 Years	3 Years	1 Year
Dukes								
Development	Dukes and Bussey (20)	St. Mark's Hospital, London, England	1928–1952	Rectal cancer	42	0.84 (0.82–0.87)	–†	–†
Validation 1	Jass et al. (21)	St. Mark's Hospital, London, England	1960–1969	Rectal cancer	29	0.78 (0.74–0.82)	–†	–†
Validation 2	Neoptolemos et al. (8)	Birmingham, England	1967–1976	Colon and rectal cancer	27	0.81 (0.69–0.93)	0.74	0.70
Validation 3	Secco et al. (41)	Genova, Italy	1978–1983	Colon and rectal cancer	43	0.73 (0.62–0.84)	0.71	0.64
Validation 4	Harrison et al. (11)	Memphis, Tennessee	1964–1983	Rectal cancer	47	0.74 (0.68–0.80)	0.73	0.71
Validation 5	Fisher et al. (17)	Canada and the United States	After 1983	Rectal cancer	46	0.76 (0.71–0.80)	0.78	0.64
Jass								
Development	Jass et al. (21)	St. Mark's Hospital,London, England	1960–1965	Rectal cancer	26	0.85 (0.80–0.90)	–†	–†
Validation 1	Jass et al. (21)	St. Mark's Hospital, London, England	1965–1969	Rectal cancer	33	0.80 (0.75–0.85)	–†	–†
Validation 2	Neoptolemos et al. (8)	Birmingham, England	1967–1976	Colon and rectal cancer	30	0.80 (0.58–1.0)	0.76	0.68
Validation 3	Secco et al. (19)	Genova, Italy	1978–1983	Colon and rectal cancer	49	0.67 (0.56–0.77)	0.66	0.67
Validation 4	Harrison et al. (11)	Memphis, Tennessee	1964–1983	Rectal cancer	47	0.79 (0.74–0.84)	0.77	0.75
Validation 5	Fisher et al. (17)	Canada and the United States	After 1983	Rectal cancer	48	0.77 (0.73–0.80)	0.77	0.70

* Numbers in parentheses are 95% CIs. Whenever possible, the parametric estimate of the area under the receiver-operating characteristic curve was used to minimize the influence of the number of stages on this area (44).

† No data were available.

Table 4. Dukes and Jass Systems for Predicting 5-Year Mortality Associated with Colon and Rectal Cancer*

System	Sample Size	Overall Mortality Rate	Mortality Rate by Stage (Distribution by Stage)						
			A	B	B1	B2	C	C1	C2
	n	%	⟵ % (%) ⟶						
Dukes revision									
Development	2037	42	2 (15)	22 (34)			68 (50)		
Validation 1	710	29	3 (16)	19 (44)			–	49 (35)	77 (4)
Validation 2	91	27	0 (5)	14 (62)			52 (32)		
Validation 3	110	43	0 (4)	–	29 (25)	36 (30)	62 (41)		
Validation 4	348	47	22 (32)	53 (30)			69 (38)		
Validation 5	687	46	18 (25)	41 (28)			–	57 (30)	79 (17)

* The Dukes system has undergone many modifications, but all used the A-B-C-D system. Stage D was excluded in this analysis because it was available in only one study. Where stages were subdivided in the study, they are reported and analyzed as divided.

among sites). In addition, multisite sampling allows some assessment of spectrum transportability because the spectrum of disease may vary by site.

4. *Multiple independent validations* test the accuracy of the system in the hands of diverse independent investigators at diverse geographic sites. They thereby provide more thorough evidence of methodologic transportability than any single validation, even if that single validation is conducted at multiple sites. Differences in the methods of patient selection and data collection between the development study and more common clinical practice are more likely to be detected when multiple independent investigators have attempted to apply the system.

5. *Multiple independent validations with varying follow-up periods* test the accuracy of the system across multiple independent investigators, geographic sites, and follow-up periods. All other things being equal, a system that can predict the outcome over many different follow-up periods is more generalizable. If detailed life tables are reported (68) rather than a single cut-point of 5 or 10 years, it is possible to calculate the discrimination and calibration of a system for any intermediate time point that might be of interest.

Applying the System to the Patient

Using the background information that you just gained, you organize data from the articles comparing the Dukes and Jass systems into three summary tables (**Tables 3**, **4**, and **5**). (For details of how these tables were generated, see the Appendix). Because neither system specified a "predicted rate" of outcomes by stage, you cannot evaluate the calibration of the system in development. However, you can compare the rates observed by stage in subsequent validations to see whether they are similar to those from the development samples (**Figures 1** and **2**). Overall, both systems fail to maintain calibration in subsequent validations. For example, the mortality rate in the average patient with Dukes stage B disease ranged from 14% to 53% (**Table 4**, **Figure 1**). In patients with Jass stage III disease, 5-year mortality rates ranged from 33% to 67% (**Table 5**, **Figure 2**). You conclude that, in general, these systems are not well calibrated.

However, both the Dukes and Jass systems have more consistent rates for this patient's more advanced stage of disease. Dukes stage C has a 5-year mortality rate of 49% to 79% (**Table 4**). (Mortality rates for stage C1 disease are modestly lower and more uniform—between 49% and 57%—but only two studies report on stage C1 disease separately.) The mortality rate associated with Jass stage IV disease ranges from 70% to 84% (**Table 5**, **Figure 2**). You conclude that the calibration of the Dukes and Jass systems is better in more advanced stages of disease. Furthermore, you see that both the Dukes and Jass staging systems were reasonably discriminating in the development samples (areas under the ROC curve, 0.84 and 0.85, respectively) and that their subsequent discrimination has varied but has been generally good. But how diverse are the settings in which these systems have been tested? Are there specific settings that pertain more to this patient's circumstances?

Reproducibility

No report of the Dukes or Jass system used resampling techniques. Therefore, you cannot determine whether these models were overfit in development. Although the simplicity of the models makes underfitting more likely, overfitting is still possible if the investigators evaluated many candidate variables before identifying those included in the model.

Historical Transportability

The development and validations of the Jass system span 23 years (1960 to 1983). The Dukes system has undergone an even more extreme test: It was developed on a patient sample from 1928 to 1952 and has been validated in samples from 1960 to 1983. Remarkably, the prognosis for colon and rectal cancer has not improved much over this time.

Nevertheless, the calibration of both staging systems has changed with time so that the studies that were conducted closer in time to the development sample (validations 1 and 2) have, in general, mortality rates by stage that are more similar to the development sample stages than to those that are less close in time (**Figures 1** and **2**). For purposes of quoting mortality rates, it seems more appropriate to emphasize the most recent data to your patient (that is, validation 5 suggests a 5-year mortality rate of 57% for Dukes stage C1 disease and 80% for Jass stage IV disease [**Tables 4** and **5**]). The discrimination of both systems seems to be less time dependent (**Table 3**).

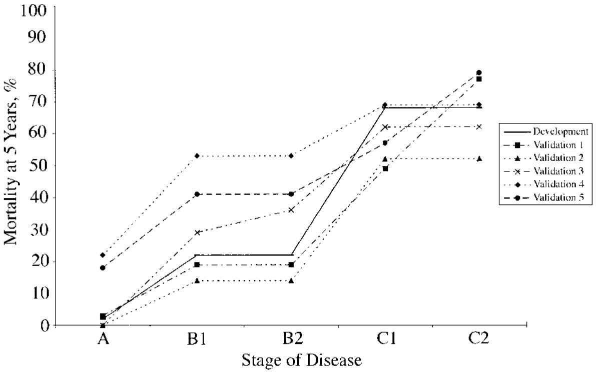

Figure 1. Calibration of the Dukes staging system. Points on lines correspond to the 5-year mortality rate by Dukes stage for each of five validation studies and the original development report (**Table 3**). To show point estimates for stage B1, B2, C1, and C2 disease when reported, point estimates for stage B and C were repeated in studies in which mortality rates were not reported by substages within stages B and C.

Geographic Transportability

The Dukes and Jass systems were developed at the same hospital in London. Subsequent validations have been done in Birmingham, England; Genoa, Italy; Memphis, Tennessee; and other sites in the United States and Canada. The discrimination of both systems seems to have reasonable geographic transportability. Fortunately, one of the more recent validations of these systems was conducted in Canada and the United States, so we need not be concerned that these systems (developed in European patients) do not apply to U.S. citizens, such as the patient from Cleveland.

Methodologic Transportability

Two major issues of methodologic transportability are addressed by these validations: whether these systems, developed to stage rectal cancer, can also be applied to colon cancer (as in validations 2 and 3) and whether the method of measuring variables used in development can be transported to other investigators in the validation studies.

Methodologic transportability to colon cancer seems to be reasonably good for the Dukes staging system. The discrimination in validations 2 and 3 for this systems seems to be similar to that in other validations in which the studies were restricted to patients with rectal cancer. In contrast, the area under the ROC curve for validation 3 of the Jass system is poor (0.67) and has degraded substantially from the development sample (area under the ROC curve, 0.85) (**Table 3**). Similarly, the calibration curve for validation 3 of the Jass system has deviated substantially from the development values (**Figure 2**). Thus, the generalizability of the Jass staging system to colon cancer appears suspect. Furthermore, methodologic transportability of this system to other investigators may also be unreliable (12). Because your patient has colon cancer and his pathologic specimen was not read at St. Mark's Hospital, you may want to rely more heavily on the Dukes staging system.

Spectrum Transportability

Overall 5-year mortality has varied widely over the validations of the Dukes and Jass systems, from a low of 26% to a high of 49%. This suggests a wide range of disease spectrum. Whereas the discrimination of Dukes stages (developed in a sample with an overall mortality rate of 42%) does not seem to be closely tied to overall mortality, discrimination in the Jass system is better in samples with a lower

Table 5. Jass System for Predicting 5-Year Mortality Associated with Colon and Rectal Cancer*

System	Sample	Overall Mortality Rate	Mortality Rate by Stage (Distribution by Stage)			
			I	II	III	IV
	n	%	←	% (%)		→
Jass						
Development	379	26	4 (31)	15 (31)	33 (18)	73 (20)
Validation 1	331	33	8 (27)	19 (33)	54 (22)	74 (17)
Validation 2	91	30	6 (16)	8 (49)	NA	70 (8)
Validation 3	121	49	16 (10)	43 (36)	54 (48)	84 (6)
Validation 4	348	47	15 (22)	29 (25)	58 (27)	79 (26)
Validation 5	722	48	22 (29)	39 (27)	67 (24)	80 (20)

* NA = not available.

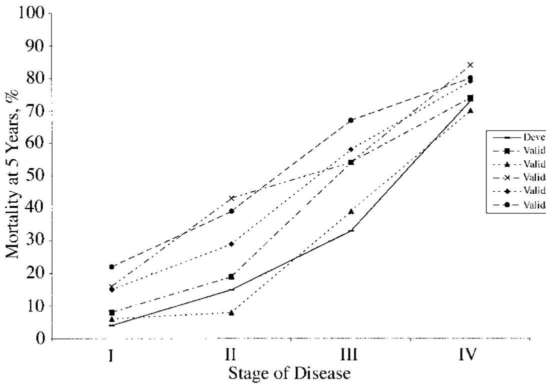

Figure 2. Calibration of the Jass staging system. Points on lines correspond to the 5-year mortality rate by Jass stage for each of five validation studies and the original (development) report (**Table 3**). Because validation 2 reported no point estimate for stage III disease, a midpoint between stage II and stage III was imputed by averaging the difference between the stages.

overall mortality rate. Of note, the Jass system was developed in a sample in which the overall mortality rate was 26% (**Table 5**). Given that your patient is from an age group that tends to have more aggressive disease (that is, greater severity), it would again seem more appropriate to emphasize use of the Dukes staging system.

Follow-up Period Transportability

Both the Dukes and Jass systems seem to discriminate better over observation periods longer than 1 year. Of note, the differences in discrimination between 1 and 3 years are more pronounced than those between 3 and 5 years (**Table 3**). It is reassuring to know that these systems provide reasonably accurate long-term information. Furthermore, the ability of both systems to discriminate over longer periods suggests that both systems are relatively robust to differences in follow-up period.

Cumulative Level of Validation

The external validity of the Dukes and Jass systems has been comprehensively tested (**Tables 3**, **4**, and **5** and **Figures 1** and **2**). These tests include an extreme test of historical transportability, a cumulative test of geographic transportability that represents several European and North American countries, tests of methodologic transportability, substantial tests of spectrum transportability, and tests of follow-up period transportability that span 5 years. These have been undertaken by multiple independent investigators in diverse locations over an extended period. Therefore, the cumulative level of tested generalizability for both systems corresponds to a level 5 (multiple independent validations with varying follow-up periods). However, these systems did not maintain calibration and discrimination in all of these tests. Thus, the level of external validity achieved is more modest.

Overall, both the Dukes and Jass systems were subject to substantial variation in calibration and discrimination. By classifying this variation by stage of disease and by the types of transportability tested, it was possible to better understand the limits of the systems' cumulative generalizability. Although neither system remained calibrated overall, both were somewhat better calibrated among patients with more advanced disease. Furthermore, the Jass system did not seem to generalize well to colon cancer or new investigators, but the Dukes system did.

Thus, we can identify useful information for your patient. Specifically, you can tell him that both the Dukes and Jass staging systems have been widely tested and proved reasonably accurate and generalizable in more advanced disease. You can tell him that both systems categorize his disease as a higher stage, one associated with a 5-year survival rate of 40% or 50% or a mortality rate of 50% to 60%. By using the life tables from these reports, you can also offer him estimates for more intermediate intervals of survival should he desire them. The way in which you impart this information deserves careful attention to presentation (69–71), the spiritual and psychological impact of the information (22), and the patient's right to know (72, 73). Other sources have addressed these issues (74–76).

Limitations

Our method of assessing the generalizability of prognostic information has several important limitations. First, the method does not account for the quality of studies; instead, it depends on how they compare with each other. Other sources offer guidelines on how to evaluate the quality of a prognostic study (25, 27–40). Nevertheless, we propose a method for evaluating the cumulative external validity of a prognostic system that we believe is a more definitive test of quality.

Second, the Dukes and Jass systems were developed before bootstrapping methods were widely used. Thus, we could not determine the degree to which subsequent failures in validation are due to lack of reproducibility (internal validity) rather than solely problems in transportability (external validity).

Third, the Dukes and Jass staging systems are simple and discontinuous and have discrete levels (3 and 4 levels, respectively). We chose these systems because we wanted to make our examples accessible to persons who are not familiar with complex prognostic models. These systems may be more susceptible than systems developed by using more sophisticated techniques to limitations in transportability. Nevertheless, the basic method of evaluation presented here can be applied to any prognostic technology that predicts a single event (for example, first hospitalization or low birth weight), and most of the methods can be extended to survival analysis with censored observations (35, 37).

Finally, we encourage physicians to use as many sources of externally validated prognostic information as they can find. The Dukes and Jass systems were chosen for illustration purposes. It is possible that other systems would offer additional helpful information to a patient like the one discussed here.

Conclusions

Generalizability, also known as external validity (41, 64), is the ability of a prognostic system to offer accurate predictions in subsequent samples of patients. Accuracy is composed of discrimination and calibration. A system that does not discriminate in subsequent studies is unlikely to remain calibrated (37, 41), but many systems lose calibration while maintaining some discrimination. Because the relative importance of calibration and discrimination depends on the intended application, it is important to test both. Furthermore, the generalizability of a predictive system depends on its reproducibility and transportability. Reproducibility can be assessed by using bootstrapping techniques in the development sample, but the assessment of transportability requires new data.

As accurate and generalizable prognostic information increases in importance to health care, methods for generating this information are becoming increasingly complex. Like any other scientific hypothesis, the external validity of a prognostic system is established by being cumulatively tested and found accurate across increasingly diverse settings. The more diverse the previous settings in which the system has been tested and found accurate, the more likely it will generalize to an untested setting. By providing an approach to evaluating the accuracy and cumulative generalizability of these systems, we aim to help clinicians and researchers become more informed users of prognostic technology.

Appendix

Candidate references were identified by doing a MEDLINE search from December 1997 back to 1966 with the text words *Jass* and *Dukes.* These combined terms generated 18 listings (1–18). Of these 18 studies, 3 reported no original data (7, 13, 16) and 4 did not include survival as an outcome (2, 4, 9, 10). Of the remaining 11 studies of survival with original data, 6 did not report data for the Dukes and Jass systems in a manner that would allow generation of a mortality table (1, 3, 5, 6, 12, 18). Another paper was published in a hard-to-find journal (*Minerva Chirurgica*) (15), but a search on the first author's name revealed that the same author had published a paper with the same title in the same year in *Digestion* (19); thus, this paper was substituted. This paper and the remaining 3 papers identified by the original search reported mortality rates by stage for both systems (8, 11, 17, 19). In addition, the references of these papers were searched, yielding original reports of both systems (20, 21).

Characteristics of the sample and study, such as setting, year of study, spectrum of disease, and follow-up period, were recorded. In addition, the actual fraction of patients experiencing the outcome of interest was collected directly or was calculated from the available information. If mortality was shown as a Kaplan–Meier plot, mortality was independently determined at 1, 3, and 5 years by two of the authors. These determinations resulted in a 98% (87 of 89 readings) agreement within 5 absolute percentage points for mortality. Because most reports did not specify when observations were censored, we used the sample size at baseline to calculate numerators and denominators from reported rates. In cases of disagreement, differences between the two readings were averaged. Essential information included the percent mortality by stage for both development and validation samples and the numerator and denominator used to calculate this percentage (that is, the number of deaths that occurred in each stage and the number of patients in each stage). The area under the ROC curve was calculated from the reported data by using software in the public domain (44).

Grant Support: This work is derived in part from a doctoral dissertation on health systems by Dr. Justice for the Graduate Group in Managerial Science and Applied Economics, University of Pennsylvania (1996), which was supported by Agency for Health Care Policy and Research National Research Service Award #5 T32 HS00009 from the University of Pennsylvania.

Requests for Reprints: Amy C. Justice, MD, PhD, Pittsburgh Veterans Affairs Health Care System, Section of General Internal Medicine (11E), University Drive C, Pittsburgh, PA 15240.

Current Author Addresses: Dr. Justice: Pittsburgh Veterans Affairs Health Care System, Section of General Internal Medicine (11E-120), University Drive C, Pittsburgh, PA 15240.
Dr. Covinsky: Division of Geriatrics, 111G, Veterans Affairs Medical Center, 4150 Clement Street, San Francisco, CA 94121.
Dr. Berlin: Center for Clinical Epidemiology and Biostatistics, University of Pennsylvania, 423 Guardian Drive, Room 815, Blockley Hall, Philadelphia, PA 19104-6021.

References

1. **Poller DN, Baxter KJ, Shepherd NA.** p53 and Rb1 protein expression: are they prognostically useful in colorectal cancer? Br J Cancer. 1997;75: 87-93.
2. **Jass JR, Ajioka Y, Allen JP, Chan YF, Cohen RJ, Nixon JM, et al.** Assessment of invasive growth pattern and lymphocytic infiltration in colorectal cancer. Histopathology. 1996;28:543-8.
3. **Secco GB, Campora E, Fardelli R, Lapertosa G, De Lucchi F, Gianquinto D, et al.** Chromogranin-A expression in neoplastic neuroendocrine cells and prognosis in colorectal cancer. Tumori. 1996;82:390-3.
4. **Liabakk NB, Talbot I, Smith RA, Wilkinson K, Balkwill F.** Matrix metal-

loprotease 2 (MMP-2) and matrix metalloprotease 9 (MMP-9) type IV collagenases in colorectal cancer. Cancer Res. 1996;56:190-6.

5. **Öfner D, Riehemann K, Maier H, Riedmann B, Nehoda H, Totsch M, et al.** Immunohistochemically detectable bcl-2 expression in colorectal carcinoma: correlation with tumour stage and patient survival. Br J Cancer. 1995;72: 981-5.
6. **Gagliardi G, Stepniewska KA, Hershman MJ, Hawley PR, Talbot IC.** New grade-related prognostic variable for rectal cancer. Br J Surg. 1995;82: 599-602.
7. **Vecchio FM.** The pathologist's role in the diagnosis and therapy of rectal cancer. Rays. 1995;20:15-20.
8. **Neoptolemos JP, Oates GD, Newbold KM, Robson AM, McConkey C, Powell J.** Cyclin/proliferation cell nuclear antigen immunohistochemistry does not improve the prognostic power of Dukes' or Jass' classifications for colorectal cancer. Br J Surg. 1995;82:184-7.
9. **Darzi A, Lewis C, Menzies-Gow N, Guillou PJ, Monson JR.** Laparoscopic abdominoperineal excision of the rectum. Surg Endosc. 1995;9:414-7.
10. **Gagliardi G, Kandemir O, Liu D, Guida M, Benvestito S, Ruers TG, et al.** Changes in E-cadherin immunoreactivity in the adenoma-carcinoma sequence of the large bowel. Virchows Arch. 1995;426:149-54.
11. **Harrison JC, Dean PJ, el-Zeky F, Vander Zwaag R.** From Dukes through Jass: pathological prognostic indicators in rectal cancer. Hum Pathol. 1994; 25:498-505.
12. **Deans GT, Heatley M, Anderson N, Patterson CC, Rowlands BJ, Parks TG, et al.** Jass' classification revisited. J Am Coll Surg. 1994;179:11-7.
13. **Fucci L, Pirrelli M, Caruso ML.** Carcinoma and synchronous hyperplastic polyps of the large bowel. Pathologica. 1994;86:371-5.
14. **Öfner D, Totsch M, Sandbichler P, Hallbrucker C, Margreiter R, Mikuz G, et al.** Silver stained nucleolar organizer region proteins (Ag-NORs) as a predictor of prognosis in colonic cancer. J Pathol. 1990;162:43-9.
15. **Secco GB, Fardelli R, Lapertosa G, Fulcheri E, Rovida S, Ratto GB, et al.** [The prognostic value of Jass' histopathological classification of cancer of the left colon and rectum.] Minerva Chir. 1990;45:1347-53.
16. **Jass JR.** Dukes and Jass systems [Letter]. Dis Colon Rectum. 1990;33:721-2.
17. **Fisher ER, Robinsky B, Sass R, Fisher B.** Relative prognostic value of the Dukes and the Jass systems in rectal cancer. Findings from the National Surgical Adjuvant Breast and Bowel Projects (Protocol R-01). Dis Colon Rectum. 1989;32:944-9.
18. **Stahle E, Enblad P, Pahlman L, Glimelius B.** Can mortality from rectal and rectosigmoid carcinoma be predicted from histopathological variables in the diagnostic biopsy? APMIS. 1989;97:513-22.
19. **Secco GB, Fardelli R, Campora E, Lapertosa G, Fulcheri E, Rovida S, et al.** Prognostic value of the Jass histopathologic classification in left colon and rectal cancer: a multivariate analysis. Digestion. 1990;47:71-80.
20. **Dukes CE, Bussey HJ.** The spread of rectal cancer and its effect on prognosis. Br J Cancer. 1958;12:309-20.
21. **Jass JR, Love SB, Northover JM.** A new prognostic classification of rectal cancer. Lancet. 1987;1:1303-6.
22. **Christakis NA.** Prognostication and Death in Medical Thought and Practice. Philadelphia: Univ of Pennsylvania Pr; 1995.
23. **Christakis NA, Iwashyna TJ.** Attitude and self-reported practice regarding prognostication in a national sample of internists. Arch Intern Med. 1998; 158:2389-95.
24. **Wagner DP, Knaus WA, Harrell FE, Zimmerman JE, Watts C.** Daily prognostic estimates for critically ill adults in intensive care units: results from a prospective, multicenter, inception cohort analysis. Crit Care Med. 1994;22: 1359-72.
25. **Braitman LE, Davidoff F.** Predicting clinical states in individual patients. Ann Intern Med. 1996;125:406-12.
26. **Charlson ME, Pompei P, Ales KL, MacKenzie CR.** A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40:373-83.
27. **Laupacis A, Wells G, Richardson S, Tugwell P.** Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. JAMA. 1994;272:234-7.
28. **Heckerling PS, Conant RC, Tape TG, Wigton RS.** Reproducibility of predictor variables from a validated clinical rule. Med Decis Making. 1992;12: 280-5.
29. **Fletcher RH, Fletcher SW, Wagner EH.** Clinical Epidemiology. The Essentials. 3d ed. Baltimore: Williams & Wilkins; 1998.
30. **Feinstein AR.** Clinical biostatistics. XIV. The purposes of prognostic stratification. Clin Pharmacol Ther. 1972;13:285-97.
31. **Feinstein AR.** Clinical Judgment. Baltimore: Williams & Wilkins; 1967.
32. **Wasson JH, Sox HC, Neff RK, Goldman L.** Clinical prediction rules. Applications and methodological standards. N Engl J Med. 1985;313:793-9.
33. **Laupacis A, Sekar N, Stiell IG.** Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA. 1997;277:488-94.
34. **Senn SJ.** Covariate imbalance and random allocation in clinical trials. Stat Med. 1989;8:467-75.
35. **Harrell FE Jr, Lee KL, Califf RM, Pryor DB, Rosati RA.** Regression modelling strategies for improved prognostic prediction. Stat Med. 1984;3:143-52.
36. **Charlson ME, Ales KL, Simon R, MacKenzie R.** Why predictive indexes perform less well in validation studies. Is it magic or methods? Arch Intern Med. 1987;147:2155-61.
37. **Harrell FE Jr, Lee KL, Mark DB.** Tutorial in biostatistics. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361-87.
38. **Baxt WG.** Application of artificial neural networks to clinical medicine. Lancet. 1995;346:1135-8.
39. **Wyatt J.** Nervous about artificial neural networks? Lancet. 1995;346:1175-7.
40. **Hart A, Wyatt J.** Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks. Med Inf (Lond). 1990;15:229-36.
41. **Ash AS, Shwartz M.** Evaluating the performance of risk-adjustment methods: dichotomous measures. In: Iezzoni LI, ed. Risk Adjustment for Measuring Health Care Outcomes. Ann Arbor, MI: Health Administration Pr; 1994:313-46.
42. **Poses RM, Cebul RD, Centor RM.** Evaluating physicians' probabilistic judgments. Med Decis Making. 1988;8:233-40.
43. **Hanley JA, McNeil BJ.** The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29-36.
44. **Centor RM.** A Visicalc program for estimating the area under a receiver operating characteristic (ROC) curve. Med Decis Making. 1985;5:139-48.
45. **Diamond GA.** What price perfection? Calibration and discrimination of clinical prediction models. J Clin Epidemiol. 1992;45:85-9.
46. **Hilden J, Habbema DF, Bjerregaard B.** The measurement of performance in probabilistic diagnosis. II. Trustworthiness of the exact values of the diagnostic probabilities. Methods Inform Med. 1978;17:227-37.
47. **Hilden J, Habbema DF, Bjerregaard B.** The measurement of performance in probabilistic diagnosis. III. Methods based on continuous functions of the diagnostic probabilities. Methods Inform Med. 1978;17:238-46.
48. **Swets JA.** Measuring the accuracy of diagnostic systems. Science. 1988;240: 1285-93.
49. **Yates JF.** External correspondence: decompositions of the mean probability score. Organizational Behavior and Human Performance. 1982;30:132-56.
50. **Lindsay MR, Ehrenberg AS.** The design of replicated studies. American Statistician. 1993;47:217-28.
51. **Mosteller F, Tukey JW.** Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley; 1977.
52. **Breiman L, Friedman JH, Olshen RA, Stone CJ.** Classification and Regression Trees. Pacific Grove, CA: Wadsworth and Brooks/Cole Advanced Books and Software; 1984.
53. **Concato J, Peduzzi P, Holford TR, Feinstein AR.** Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol. 1995;48:1495-501.
54. **Justice AC, Feinstein AR, Wells CK.** A new prognostic staging system for the acquired immunodeficiency syndrome. N Engl J Med. 1989;320:1388-93.
55. **Justice AC, Aiken LH, Smith HL, Turner BJ.** The role of functional status in predicting inpatient mortality with AIDS: a comparison with current predictors. J Clin Epidemiol. 1996;49:193-201.
56. **Stone VE, Seage GR 3d, Hertz T, Epstein AM.** The relation between hospital experience and mortality for patients with AIDS. JAMA. 1992;268: 2655-61.
57. **Mocroft AJ, Johnson MA, Sabin CA, Lipman M, Elford J, Emery V, et al.** Staging system for clinical AIDS patients. Royal Free/Chelsea and Westminster Hospitals Collaborative Group. Lancet. 1995;346:12-7.
58. **Cozzi Lepri A, Pezzotti P, Phillips AN, Petrucci A, Rezza G.** Clinical staging system for AIDS patients [Letter]. Lancet. 1995;346:1103.
59. **Deyo RA, Cherkin DC, Ciol MA.** Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol. 1992;45: 613-9.
60. **Ransohoff DF, Feinstein AR.** Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299:926-30.
61. **Shea JA, Berlin JA, Escarce JJ, Clarke JR, Kinosian BP, Cabana MD, et al.** Revised estimates of diagnostic test sensitivity and specificity in suspected biliary tract disease. Arch Intern Med. 1994;154:2573-81.
62. **Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS.** Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med. 1992; 117:135-40.
63. **Rothman KJ.** Modern Epidemiology. Boston: Little, Brown; 1986.
64. **Knaus WA, Draper EA, Wagner DP, Zimmerman JE.** APACHE II: a severity of disease classification system. Crit Care Med. 1985;13:818-29.
65. **Brewster AC, Karlin BG, Hyde LA, Jacobs CM, Bradbury RC, Chae YM.** MEDISGPS: a clinically based approach to classifying hospital patients at admission. Inquiry. 1985;22:377-87.
66. **Iezzoni LI, Moskowitz MA.** A clinical assessment of MedisGroups. JAMA. 1988;260:3159-63.
67. **Alemi F, Rice J, Hankins R.** Predicting in-hospital survival of myocardial infarction. A comparative study of various severity measures. Med Care. 1990;28:762-75.
68. **Lang TA, Secic M.** Assessing time to an event as an endpoint. In: How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors, and Reviewers. Philadelphia: American Coll of Physicians; 1997:137-46.
69. **O'Connor AM.** Effects of framing and level of probability on patients' preferences for cancer chemotherapy. J Clin Epidemiol. 1989;42:119-26.
70. **O'Connor AM, Pennie RA, Dales RE.** Framing effects on expectations, decisions, and side effects experienced: the case of influenza immunization. J Clin Epidemiol. 1996;49:1271-6.
71. **Mazur DJ, Merz JF.** How the manner of presentation of data influences older patients in determining their treatment preferences. J Am Geriatr Soc. 1993;41:223-8.
72. **Annas GJ.** Informed consent, cancer, and truth in prognosis. N Engl J Med. 1994;330:223-5.
73. **Smith TJ, Swisher K.** Telling the truth about terminal cancer [Editorial]. JAMA. 1998;279:1746-8.
74. **Buckman R, Kason Y.** How to Break Bad News: A Guide for Health Care Professionals. Baltimore: Johns Hopkins Univ Pr; 1992.
75. **Brewin TB.** Three ways of giving bad news. Lancet. 1991;337:1207-9.
76. **Ptacek JT, Eberhardt TL.** Breaking bad news. A review of the literature. JAMA. 1996;276:496-502.