Normal Tissue Effects: Reporting and Analysis

Soren M. Bentzen, Wolfgang Dörr, Mitchell S. Anscher, James W. Denham, Martin Hauer-Jensen, Lawrence B. Marks, and Jacqueline Williams

Any effective cancer therapy developed to date is associated with a spectrum of normal tissue effects of varying incidence and severity. With an increasing number of novel therapeutic approaches undergoing clinical testing and an increased effort to optimize the established treatment modalities, methods for reliable quantification of normal tissue effects have become a key element in advancing cancer care. Here, we present a review of many of the issues involved in reporting and analyzing clinical normal tissue effect data. A distinction is introduced between explorative (science-driven) and pragmatic (patient-centered) studies. The desirable properties of criteria for reporting and grading toxicity are discussed from a biological and clinical perspective. Validation of toxicity criteria and the statistical issues involved in analyzing this type of data are presented with special emphasis on descriptors of the time evolution of toxicity. Finally, we discuss surrogate markers for late effects, mechanistic studies, and the design of clinical studies with normal tissue endpoints as a primary outcome. It is concluded that a consensus is required on guidelines for the reporting of normal tissue effects to improve the comparability of published reports on treatment outcome.

© 2003 Elsevier Inc. All rights reserved.

Conventional cancer therapy is based on 3 treatment modalities: surgery, radiotherapy, and chemotherapy alone or in combination. Surgical toxicity is mainly expressed by local scarring of the normal tissue. Conversely, systemic chemotherapy affects both malignant and normal cells throughout the body. Radiotherapy primarily affects tissues in the vicinity of the target, but significant volumes of normal tissue in the beam's path, remote from the target, may be affected.

Cancer therapies are associated with a spectrum of normal tissue effects. Thus, an evaluation of their efficacy should contain information on both tumor and normal tissue effects. Although the reporting of tumor outcome is fairly standardized, there is no general consensus on how best to quantify normal tissue effects. The acceptable frequency and severity of these side effects obviously depends on the expected benefits from the therapy (ie, the risk/benefit ratio) but also on a number of individual factors. In this article, we distinguish between normal tissue effects of cancer therapy and treatment-related morbidity. The latter term refers to normal tissue effects that negatively impact on quality of life (QOL) or physical functioning. All tissues will express a response to therapeutic radiation doses at the molecular, histological, or clinical level. However, depending on dose, dose per fraction, volume, or other factors, not all patients exhibit symptoms or other clinical manifestations. Morbidity is therefore a subset of normal tissue effects.

The focus of this article is on radiation therapy, but much of the following applies to surgery and chemotherapy as well. There are major similarities between the early effects of radio- and chemotherapy. Late effects after radiotherapy have been studied for a century but have been considered only recently for the other modalities.

Advances in radiation technology and dose-fractionation have changed the pattern of side effects. Historically, early skin reactions were generally dose limiting because of the depth-dose profile for orthovoltage radiotherapy. Today, much of the focus has changed to late effects. Although current radiotherapy approaches generally succeed in preventing severe morphological and structural changes, such as fistulae and chronic ulcerations, insidious chronic changes in organ function (eg, of the heart, lung, or kidney) may become clinically manifest years after treatment. This picture may change again with the...
recent and expected advances in radiation and cancer biology, functional imaging, and treatment planning and delivery that may facilitate high local doses.

Routine Versus Toxicity-Specific Investigations

The study of normal tissue side effects has 3 main aims: (1) to serve as an integral part of quality assurance in routine practice, including ongoing management of morbidity; (2) to establish the type, incidence and severity of effects for specific therapies to inform the decision making by patients, physicians, and health care managers; and (3) to investigate the pathobiology underlying these effects, and thereby develop strategies for their prevention or amelioration. These are toxicity-specific studies with an in-depth assessment of one or more toxicity items as primary study endpoints. This may improve the QOL of cancer survivors and/or allow an intensification of therapy to improve cure rates.

Corresponding to these aims, there are 2 trends in normal tissue effects research, one being pragmatic and patient centered and the other being mechanistic and biology centered. Table 1 summarizes a number of methodological aspects that tend to differ or at least are emphasized to a varying extent with the 2 types of studies. Clearly, careful clinical observation and documentation of side effects, early as well as late, is required in any case. The data obtained should ideally be reported in a format that facilitates comparison between studies and allow analysis by other researchers. This requires comprehensive, validated, and widely accepted scoring systems. Arguably, the failure to devise such a system is an impediment to normal tissue effects research. On the other hand, it is not obvious that a single system could be devised that would be equally useful for the various research aims.

Also, establishing a set of toxicity criteria will not suffice. There are a number of outstanding issues in the design and analysis of normal tissue studies that need clarification to improve their scientific value, and these will be discussed further.

Distinguishing Early Versus Late Effects

From a biological and clinical point of view, it is useful to have a nominal division of normal tissue effects into early and late effects. Early effects are expressed during or immediately after the end of therapy, whereas late effects may become manifest after latent periods of months to years. Early reactions are usually reversible and therefore often considered less relevant for limiting treatment intensity. However, a recent review found that there may be an association between the occurrence of early reactions and the risk of

---

Table 1. Typical Characteristics of Pragmatic and Explorative Scoring Systems

<table>
<thead>
<tr>
<th>Routine Studies</th>
<th>Toxicity-Specific Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple endpoints often merged into a single grade</td>
<td>The emphasis is on a specific endpoint, varying grade</td>
</tr>
<tr>
<td>Easy to assess clinically</td>
<td>May require specific diagnostic procedures</td>
</tr>
<tr>
<td>Complete: any clinical occurrence of a late effect should be gradable and recordable using the dictionary</td>
<td>Selective: the focus is on a specific biological effect of therapy that may be quantified and subjected to mechanistic studies</td>
</tr>
<tr>
<td>Symptoms and QoL: these endpoints are most important because of their relevance to the patient.</td>
<td>Analytic and objective signs are the preferred endpoints as these can be validated across centres/observers</td>
</tr>
<tr>
<td>The focus is often on severely debilitating toxicity</td>
<td>Less severe toxicity is of interest as the increased incidence improves the statistical power of a study</td>
</tr>
<tr>
<td>Statistical methods: Kaplan-Meier, Prevalence, Cumulative incidence</td>
<td>Statistical method: Kaplan-Meier</td>
</tr>
<tr>
<td>Follow patients for their lifetime</td>
<td>Follow patients long enough to ensure statistical power</td>
</tr>
<tr>
<td>Endpoints selected based on clinical relevance</td>
<td>Endpoints are selected to provide high sensitivity and high specificity</td>
</tr>
<tr>
<td>Focus on the patient</td>
<td>Focus on biology</td>
</tr>
</tbody>
</table>
subsequent late effects in a number of organs.\textsuperscript{1} In contrast, late effects are generally considered irreversible and progressive. Denham and colleagues\textsuperscript{2} recently proposed to replace the simple early/late classification of normal tissue effects with a more mechanistically based classification. Although this change of paradigm seems highly desirable in terms of advancing the development of strategies for prevention and treatment of normal tissue effects, it may be premature to discard the distinction between early and late effects when reporting toxicity.

There is no consensus in the literature concerning the exact definition of early and late effects. Often an operational definition is used, simply classifying effects based on an arbitrary cutoff for their latent period. One cutoff that has been used is 90 days after the onset of the treatment. It has, however, been proposed to define late effects as those that occur or have not healed by 90 days after the end of therapy.\textsuperscript{3} This latter definition seems more appropriate with many combined modality therapies extending the period of active therapy over several months. Unfortunately, there is no general biological justification for any of these conventions. However, a classification of early and late effects is clinically relevant, and therefore it is essential for comparability between studies that a common definition is agreed on. Recently, new guidelines have been proposed for clarifying early versus late effects based on aggregate data review. The application and precise methods for this have yet to be fully described. Ideally, it would be desirable if this definition could be endpoint specific based on a biological consideration of the actual time course of effects for a given normal tissue effect.

**Recording of Early Effects**

As noted, early effects of cancer therapy usually occur during or shortly after the treatment (ie, at a time when the patient is still under close observation by the treating physician). Moreover, early effects (eg, in skin) have been used as a biological dosimeter in the early decades of radiotherapy; just like hematological toxicity has been dose limiting for many cytotoxic drugs. Thus, there is a fairly comprehensive literature on early morbidity after both radiotherapy and chemotherapy. This has resulted in the development and implementation of scoring systems designed for radiotherapy, like the Radiation Therapy Oncology Group/European Organisation for Research and Treatment of Cancer system (RTOG/EORTC), and chemotherapy, like the Common Toxicity Criteria or the World Health Organization system. Recently, these have been amalgamated in the Common Toxicity Criteria version 2.0,\textsuperscript{4} which is aimed at creating a common system applicable for each of the 2 modalities alone or in combination. Although developed specifically for reporting early effects in clinical trials, they have also been widely applied in routine practice.

More elaborate scoring systems have been used in clinical studies focusing on a specific early side effect. Some of these have been extensions or modifications of the RTOG/EORTC system, like for example the inclusion of the affected area in the evaluation of oral mucositis\textsuperscript{5}; others have increased the level of detail on the more functional, subjective consequences of the reaction for the patient, like the morbidity scoring system by Dische\textsuperscript{6} (which is available as a draft but unfortunately has never been published in its entirety). Clearly, the use of more in-depth toxicity scoring instruments will typically increase the time spent on each follow-up of a patient, but this may be justified by the additional information obtained.

**Recording of Late Effects**

Toxicity criteria for late effects are much less established than those for early effects. The most ambitious attempt so far to develop a comprehensive system for the grading and recording of late radiation effects was the Subjective, Objective, Management, Analytic/Late Effects Normal Tissue (SOMA/LENT) system, first published in 1995.\textsuperscript{7,8} We will herein consider the features of this system in some detail as an example of some of the general issues involved in late morbidity recording. The SOMA system was devised in the early 1990s, when working parties formed by the RTOG in the United States and the EORTC joined forces in an ambitious attempt to devise a rational, comprehensive scoring system for use in radiation oncology. As suggested by its acronym, the characteristic feature of the SOMA/LENT system is that each toxicity item is classified by
its subjective symptoms, objective signs, and management-related or analytical measures.

Table 2 shows the toxicity items for lung in the SOMA/LENT system. Subjective symptoms for this organ are cough, dyspnea, and chest pain/discomfort. A patient presenting with occasional cough and who is breathless on mild exertion would thus have a grade 1 cough and a grade 2 dyspnea. Objective signs would be radiologic changes on a chest radiograph or reduced lung function. The Management would depend on the symptom but includes prescription of narcotic or nonnarcotic drugs, supplemental oxygen, corticosteroids, respirator, or surgical interventions. In addition to the S, O, M components, a number of specific analytic tests are suggested. In case of lung function, some test outcomes are graded (eg, pulmonary functional tests), whereas others are not (eg, pulmonary lavage and computed tomography/magnetic resonance imaging scans).

### Table 2. Example of LENT/SOMA scale: Toxicity Items for the Lung

<table>
<thead>
<tr>
<th>Subjective</th>
<th>Grade 1</th>
<th>Grade 2</th>
<th>Grade 3</th>
<th>Grade 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cough</td>
<td>Occasional</td>
<td>Intermittent</td>
<td>Persistent</td>
<td>Refractory</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>Breathless on intense exertion</td>
<td>Breathless on mild exertion</td>
<td>Breathless at rest, limits all activities</td>
<td>Prevents any physical activity</td>
</tr>
<tr>
<td>Chest pain/discomfort</td>
<td>Occasional &amp; minimal</td>
<td>Intermittent &amp; tolerable</td>
<td>Persistent &amp; intense</td>
<td>Refractory &amp; excruciating</td>
</tr>
<tr>
<td>Objective</td>
<td>Pulmonary fibrosis</td>
<td>Radiological abnormality</td>
<td>Patchy dense abnormalities on radiograph</td>
<td>Dense confluent radiographic changes limited to radiation field</td>
</tr>
<tr>
<td>Lung function</td>
<td>10%-25% reduction of respiration volume and/or diffusion capacity</td>
<td>&gt;25%-50% reduction of respiration volume and/or diffusion capacity</td>
<td>&gt;50%-75% reduction of respiration volume and/or diffusion capacity</td>
<td>&gt;75% reduction of respiration volume and/or diffusion capacity</td>
</tr>
<tr>
<td>Management</td>
<td>Pain</td>
<td>Occasional non-narcotic</td>
<td>Regular non-narcotic</td>
<td>Regular narcotic</td>
</tr>
<tr>
<td>Cough</td>
<td>Non-narcotic</td>
<td>Occasional O₂</td>
<td>Continuous O₂</td>
<td>Surgical intervention</td>
</tr>
<tr>
<td>Analytic</td>
<td>PFT</td>
<td>Decrease to &gt;75%-90% of preTx value</td>
<td>Decrease to &gt;50%-75% of preTx value</td>
<td>Decrease to &gt;25%-50% of preTx value</td>
</tr>
<tr>
<td>DLCO</td>
<td>Decrease to &gt;75%-90% of preTx value</td>
<td>Decrease to &gt;50%-75% of preTx value</td>
<td>Decrease to &gt;25%-50% of preTx value</td>
<td></td>
</tr>
<tr>
<td>% O₂/CO₂</td>
<td>saturation</td>
<td>Decrease to &gt;70% O₂, ≤50% CO₂</td>
<td>Decrease to &gt;60% O₂, ≤60% CO₂</td>
<td></td>
</tr>
<tr>
<td>CT/MRI</td>
<td>Assessment of lung volume and zones of fibrosis</td>
<td>Assessment of pulmonary blood flow and alveolar filling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Perfusion scan</td>
<td>Assessment of cells and cytokines</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 1 shows in schematic form the trade-off between specificity and patient relevance of various dimensions of normal tissue effects.

Figure 1 Schematic representation of the trade-off between specificity and patient relevance of various dimensions of normal tissue effects.

Ranking and Classifying the Severity of Adverse Effects

Most clinical manifestations of normal tissue effects of radiotherapy may be ranked according to their severity, and this grading (also called scoring) should ideally be the subject of validation studies. One example is late bladder toxicity assessed by urination frequency in which grade 1 to 4 are defined as 3- to 4-hour intervals, 2- to 3-hour intervals, 1- to 2-hour intervals, and hourly, respectively. Increasing grades are defined so that they correspond to increasing levels of biological effect and/or increasing detriment to the QOL or physical functioning. Statistically, a graded-scale endpoint is also referred to as an ordinal variable. No assumptions are required in regards to a numerical relationship between the different grades (e.g., a grade 4 reaction is not in any sense twice as bad as a grade 2); the only assumption is that an ordering of the categories is possible and meaningful.

Several reported studies dichotomize graded endpoints into, say, none-mild and moderate or severe reactions. This procedure is evidently associated with a loss of information and should generally be avoided unless there are clinical or biological arguments for such dichotomization. Simulation studies have shown that the number of patients required to achieve a certain statistical precision of an estimate of a radiobiological parameter in a clinical study may be up to 3 times higher using a dichotomous endpoint rather than a graded scale. Graded response data for late effects may be analyzed using a generalization of the Cox model or the mixture model. For graded endpoints without censoring, ordinal logistic regression may be used.

Similarly, a cystometric assessment of bladder volume will effectively produce a continuum of reduced bladder capacity after radiotherapy. Again, a continuous scale endpoint may be reduced—at the price of a loss of information—to an ordinal endpoint by defining a limited number of grades corresponding to varying ranges of bladder volume.

Although the reduction of continuous scale to ordinal and again to dichotomous endpoints is associated with a loss of information, this may convey some advantages in terms of computation or in terms of the interpretation of the findings of a study. Most importantly, binary data form the basis for the analysis of radiation dose-response relationships. The typical sigmoid dose-response curves are, more precisely, dose-incidence curves representing the proportion of patients who have reached a certain threshold response as a func-
tion of dose. From the dose-response curve, the
effective dose (ED) after which the endpoint is
reached in, say, 50%, 5%, or 3% of the patients
(ie, the ED\textsubscript{50}, ED\textsubscript{5}, or ED\textsubscript{3}) can be estimated
together with an estimate of the statistical un-
certainty in these quantities. The steepness of
the dose-response curve is most often quantified
by the $\frac{\gamma}{\gamma_{50}}$ value (ie, the normalized dose-response
gradient at the steepest part of the curve). For a
standard logistic dose-response curve, this is at
the 50% response level, and therefore the steep-
ness of the curve is specified as $\gamma_{50}$.

**Summary Reporting of Late Effects**

From a pragmatic point of view, there have been
tries to combine the information about all
the many manifestations of treatment effect into
a single figure, an overall toxicity grade. Propon-
ents of this idea made the analogy with the
grouping of T, N, and M categories for human
tumors into clinical stages. There is, however, an
important difference between these 2 situations.
Clinical tumor stage groupings are defined based
on similarities in cause-specific survival. Using
the terminology of numerical taxonomy, cause-
specific survival is the “operational taxonomic
unit” (OTU) that will allow deciding if, say, a
T\textsubscript{3}N\textsubscript{0} and a T\textsubscript{2}N\textsubscript{1} tumor are similar (ie, whether
they should be grouped in the same clinical
stage). In the case of toxicity grading, there is no
immediately obvious OTU. One possible OTU
would appear to be the patient’s QOL as assessed
by a validated instrument. In other words, spe-
cific signs and symptoms would be grouped to-
gether if they had a similar effect on quality of
life. However, such a system would mix the actual
manifestations of normal tissue damage with how
well the patient copes with these.

In the original publication of the LENT/
SOMA system, it was suggested to add the indi-
vidual item scores for an organ and then calcu-
late the average score. This is clearly not a very
good idea. In the case of lung reactions (Table 2),
it was proposed to add the 8 items in the SOM
part of the scale and divide by 8. Thus, a patient
dying (grade 5 per definition) from restricted
pulmonary function without having recorded any
of the other 7 signs or symptoms or management
interventions for pulmonary injury would get a
score of $\frac{5}{8}=0.625$ (ie, less than a patient who
had grade 1 signs and symptoms for all the items). This flaw of the original suggestion for
using the SOMA/LENT system was almost im-
mediately recognized, and the suggestion was
withdrawn in a remarkable double publication of
letters to the editors of the *International Journal of
Radiation Oncology Biology Physics and Radiotherapy
and Oncology*. Interestingly, the original sug-
gestion has never been replaced by a more defin-
itive suggestion from the groups formulating the
SOMA system. One obvious idea is to record the
maximum grade of any toxicity item for a specific
organ/tissue and use this as the grade of toxic-
ity. Thus, a patient experiencing, say, a grade 3
loss of sphincter control after radiotherapy will
be recorded as having experienced a grade 3
rectal complication. Although this suggestion ap-
ppears to show more common sense than the orig-
inal SOMA/LENT proposal, it still involves a
number of assumptions regarding the compara-

**Validation of Toxicity Criteria**

Ideally, a proposed set of toxicity criteria should
undergo a systematic and scientific study of their
feasibility, reliability, validity, responsiveness and
specificity for the treatment before it is applied in
clinical practice or in clinical research. This, how-
ever, is a major undertaking, and it has been
largely neglected for the systems developed so
far. Each aspect of validity has several subcom-
ponents that should be addressed, and some of
these are listed in Table 3. Some elements can
only be tested in a well-designed, prospective
study. Others are soft qualities that may be dif-
ficult to quantify rigorously but nevertheless are
important for widespread acceptance of a scoring
system. Many of the concepts involved have been
developed in social sciences and psychology, and
there is rich literature on these topics. In cancer
outcomes research, most of the literature in this
area is concerned with the validation of quality of
life instruments. There is some variability in the
terminology used, but we have tried to focus on a
number of elements that would be of particular importance in evaluating toxicity criteria.

Feasibility refers to the practicality and cost of using the scale as well as its acceptability to patients and physicians. There are several examples of scoring systems that never came into widespread use because they, rightly or wrongly, were thought to be impractical, complicated to use, or excessively time consuming.

Reliability is the ability to reproduce the same toxicity scores within or between subjects or observers (Table 3). There are 3 types of reliability used in validation of grading instruments. Test-retest designs are used to determine if responses are consistent within the same observer or subject. Interobserver reliability is a critical feature of measuring consistency in grading between 2 observers. Obviously, blinding is essential in re-

### Table 3. Elements of Assessing the Performance of a Toxicity Scale

<table>
<thead>
<tr>
<th>Element</th>
<th>Subelement</th>
<th>Comments, Typical Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feasibility</td>
<td>Compliance</td>
<td>The proportion of toxicity items that are assessed (and judged to have a valid score) at a single follow-up or at a number of consecutive follow-ups</td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td>The cost in monetary and/or time units of completing a toxicity assessment for a given patient</td>
</tr>
<tr>
<td></td>
<td>Convenience</td>
<td>The ease of using a toxicity scale and the convenience to the patient/rater. This includes consideration of invasiveness, any health risks associated with the use of diagnostic ionising radiation, pain and psychological distress</td>
</tr>
<tr>
<td>Reliability</td>
<td>Intra-rater reproducibility</td>
<td>The ability of a single observer to reproduce the toxicity items and grades when presented with the same clinical or analytical information, e.g. a laboratory test or a photograph of breast appearance</td>
</tr>
<tr>
<td></td>
<td>Inter-rater concordance</td>
<td>The concordance between the toxicity grades evaluated by multiple raters when presented with the same information - this should be tested within departments, between departments and between countries</td>
</tr>
<tr>
<td></td>
<td>Test-retest reliability</td>
<td>The reliability of toxicity grades when re-testing the same patient with a minimal time interval between tests. Again, intra- and inter-rater reliability is of interest.</td>
</tr>
<tr>
<td></td>
<td>Temporal stability</td>
<td>The reliability of toxicity grades when re-testing the same patient with a specified time interval between tests. Some physiological toxicity assays may vary with time of the day or from day to day</td>
</tr>
<tr>
<td></td>
<td>Internal consistency</td>
<td>Items relating to the same aspect of toxicity should produce consistent results. Internal consistency may be checked by deliberate including repeat or slightly modified items on the scale</td>
</tr>
<tr>
<td>Validity</td>
<td>Face validity</td>
<td>Clinical relevance, common sense, biological and pathophysiological relevance</td>
</tr>
<tr>
<td></td>
<td>Content validity</td>
<td>The ability of the scale to cover the range of morbidities seen in a population of patients after a specific therapy</td>
</tr>
<tr>
<td></td>
<td>Criterion validity</td>
<td>The correlation between the toxicity item and some criterion variable of interest. An example could be the correlation between a dysphagia grade and weight loss or between a lung toxicity measure and the persons actual physical activity</td>
</tr>
<tr>
<td></td>
<td>Convergent validity</td>
<td>The correlation between a toxicity item and other assessments of the same aspect of toxicity using previously validated scales</td>
</tr>
<tr>
<td>Responsiveness</td>
<td>Sensitivity to treatment</td>
<td>The ability to detect a dose-response relationship or changes in irradiated volume, or the addition of chemotherapy or surgery. Sensitivity to interventions for morbidity.</td>
</tr>
<tr>
<td></td>
<td>intensity modification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Threshold sensitivity</td>
<td>The minimal change in the patient’s health status that can be detected on the scale.</td>
</tr>
</tbody>
</table>
liability studies. Lastly, internal consistency measures various dimensions relating to the same aspect of toxicity. Reliability is often quantified by a statistical measure of concordance such as Cronbach's alpha.

Validity refers to whether the scale measures what it is supposed to measure. Face and content validity (see Table 3) may be difficult to quantify but will typically be tested through some kind of consensus development process. Widespread practical use of a system will indirectly depend on, and help in establishing, its face and content validity. Criterion validity is the correlation between a toxicity item and some criterion variable of interest. For example, the peak grade of mucositis after head and neck radiotherapy has been shown to correlate with functional dysphagia and with the requirement for prescribed analgesics. Somewhat related to this is convergent validity, which is correlation between a particular toxicity item and the same aspect of toxicity assessed on a previously validated scale. Both criterion and convergent validity should be tested in well-designed prospective studies with an estimation of the sample size required to test a specific hypothesis concerning the strength of the association between 2 scales or between a toxicity item and a criterion variable.

Responsiveness measures sensitivity for detecting true clinical change over time and is an important aspect of any toxicity criteria. In radiation oncology, this may be interpreted as the ability of a given toxicity item to detect a radiation dose-response relationship, but it might also be established as the ability to reflect any other treatment intensification or even an intervention to relieve morbidity. Again, for a toxicity item to be of practical value, this item should exhibit responsiveness over the clinically relevant range of biologically effective doses.

Another important consideration for a given endpoint is the specificity for treatment. For example, many functional tests may be affected by the patient’s age or by intercurrent disease. This does not in itself make these endpoints invalid or even less interesting. Often, specificity may be enhanced by analyzing the change relative to a pretreatment baseline rather than the absolute measure. There are, however, statistical difficulties in quantifying the impact of toxicities having high baseline prevalence. For example, even if they are important clinically, it is very difficult to estimate the impact of accelerated coronary atherosclerosis after irradiation to the heart or of accelerated, malabsorption-related osteoporosis after radiotherapy of the abdomen or pelvis.

Very few prospective, rationally designed validation studies have been conducted on normal tissue effects scales. Most studies, allegedly addressing validity, have in reality considered feasibility or, in a few cases, convergent validity. Reliability is rarely systematically tested. Responsiveness has in some cases been established as a byproduct of studies designed with a therapeutic aim rather than in a prospective validation study. If radioresponsiveness is established from a pooling of patients with different histology or different stage of disease, special care should be taken to consider whether interpatient differences might confound the analysis.

Analyzing and Reporting the Time Evolution of Normal Tissue Effects

Early Effects

Most early effects of cancer therapy show a dynamic pattern of onset and resolution over time. After large single doses, the time course of the expression of injury is dominated by the tissue biology, and treatment-related parameters are almost irrelevant. For fractionated therapy, however, the time of occurrence of the maximum grade of reaction and the time for this to resolve will depend on the intensity of the treatment schedule as well. This raises the question of the optimal frequency of follow-ups required to get the information necessary for a sensible analysis. As an example, for specific studies of oral mucositis Maciejewski (oral communication, September 2000) has suggested that daily observations are desirable. However, once or twice weekly scoring is generally considered sufficient for this as well as for the majority of early effects. There are pragmatic issues here too. In a busy clinic, more frequent observations are not practical.

Another problem relates to the question of what is the optimal analysis and reporting of early effects data. Several descriptors have been used in the literature, and again the lack of a consensus hampers the direct comparison between studies. Among the suggestions to overcome this problem are using the average score...
over a given time period or even the entire area under the response-time curve. From both an interpretational and a statistical point of view, there is a difficulty with calculating averages of ordinal parameters. The problem is that the average is typically not defined on the scale and that taking the average involves the implicit assumption that 2 patients with a grade 2 reaction equal 1 patient with a grade 4 reaction. (The analogy would be to characterize a patient population with the average tumor stage). Admittedly, there are situations in which the area under the response-time curve may serve as a useful summary of morbidity burden. However, the usefulness of such a description—and the potential for misleading conclusions—should be carefully considered in each case.

For routine reporting, the prevalence of a specific grade of morbidity as a function of time is easier to interpret. The following terminology has been proposed. The peak prevalence is simply the maximum proportion of patients presenting with that grade of reaction at any specific follow-up. The incidence of the reaction, in contrast, is the proportion of patients who develop this grade of reaction at any point in time. As a supplement to these quantities, the time over which a defined score is experienced may be a useful descriptor. Note, that for this parameter taking averages or differences between treatment groups is immediately meaningful.

**Late Effects**

Longitudinal studies of the temporal evolution of late effects of cancer therapy have shown a wide range of latent times between irradiation and the occurrence of a specific endpoint. Telangiectasia of the skin is a classical example in which both the cumulative proportion of patients who have experienced a specific grade of reaction is increasing with time and also the severity or grade of reaction will progress over time in many individuals.

With the more widespread use of survival (or actuarial) statistics in cancer outcome studies, it was realized that quantification of the occurrence of any endpoint requiring prolonged observation of the patient requires some adjustment for the actual number of patients at risk (ie, still under observation) as a function of time after therapy. In the case of late effects after radiotherapy, the use of actuarial statistics was advocated by Hatlevoll et al and Abratt in the early 1980s.

Cumulative incidence estimates have been proposed as an alternative to the Kaplan-Meier method for quantifying morbidity as a function of time. The problem is, however, that these estimates are not specific to morbidity. This means that the cumulative incidence of morbidity after strictly identical treatments will be higher in a group of patients with a good prognosis than in an otherwise comparable group with a poor prognosis.

Prevalence estimates as a function of time have also been shown to provide useful information on the burden of late morbidity after radiotherapy. These estimates represent the proportion of patients having a specific type and grade of morbidity among all patients who are still alive and under observation at that point in time.

The relative merits of these estimates as a means of quantifying morbidity have been the topic of some discussion. In reality, these methods are not alternative but rather complementary descriptors of the time evolution of morbidity. To improve the comparability of reports, actuarial estimates should be calculated and reported as a minimum requirement. As a supplement, especially in pragmatic studies (Table 1), it may be useful to estimate cumulative incidence or prevalence of morbidity. Clearly, prevalence is an important measure of the burden of chronic toxicity on patients, their families, the health care system, and the society. Much more research is needed on the long-term evolution of morbidity in cancer survivors.

Several recent studies have been concerned with the incidence of late effects after 10 to 20 years of follow-up. In a recent paper, Jung and colleagues presented the results of a quantitative analysis of published data on the time-incidence relationship for a number of late endpoints. Jung et al proposed a simple classification of these relationships into 4 categories and observed that the observations from many published studies seemed to be well fitted by an exponential latent-time distribution. This would imply that expression of injury is a random process with a constant probability per unit time of expressing the injury among patients who have not already done so by that time (ie, a constant hazard rate). At first glance, this might seem to
indicate that very long follow-up times would be required to obtain a precise estimate of the toxicity of a given therapy. From a statistical point of view, this is actually not the case, the reason being that if the latent-time distribution is exponential, most of the events (carrying most of the information) will have occurred within the first few half times of expression.\textsuperscript{31} There are also good data sets from experimental animals suggesting that the hazard rate is not generally constant for a variety of endpoints. Ironically, this may actually lead to a need for very long observation times anyway. This would depend on the actual shape and position of the latent-time distribution for the endpoint of interest. Some endpoints, most notably treatment-induced cancer, require long follow-up to be reliably quantified.

**Adverse Effects Analysis in Oncology Clinical Trials**

Traditionally, clinical trials of cytotoxic drugs are classified as belonging to 1 of 3 phases. A phase I trial is the first testing in humans of a new drug aimed to establish the maximum acceptable dose of the drug to be taken into phase II trials. Phase II trials screen new compounds or combinations of compounds for biological effect, most often using tumour volume shrinkage (ie, clinical response) as the endpoint. Such trials are typically designed to allow early trial termination if it becomes apparent that a clinically relevant target efficacy is unlikely to be reached. Drugs that show promising response rates in phase II trials are then considered for testing in randomized controlled phase III trials aimed to test therapeutic combinations including the new compound against current best standard therapy.

This classification scheme and the prototypical study design associated with the 3 phases of trials do not work well for radiation therapy trials. Phase I studies of cytostatic drugs typically involve relatively rapid escalation of drug doses, treating for example 3 patients at each dose level before deciding whether to escalate dose. Late chemoinduced effects are not considered. In radiotherapy trials, this may not be feasible, simply because late morbidity is generally regarded as intensity limiting and this is expressed after a long latent period. Even if the study is focussing on early morbidity, there is generally not a specific dose-limiting toxicity after radiotherapy. Instead, to judge whether a modified radiotherapy schedule is clinically acceptable requires a precise estimate of incidence and severity of early effects that again means that the sample size will need to be large. Phase II trials of cytotoxic drugs use tumor volume regression as the primary endpoint. With radiotherapy, most tumors show partial or complete regression and it is therefore ultimate local tumour control that is the most important endpoint for tumour effect. Rather than sequential phase I and II studies, most modified radiotherapy schedules are piloted in feasibility studies, often comprising as many as between 50 and 100 patients when the primary endpoints are early and late morbidity and locoregional tumor control.

The design of randomized controlled phase III trials is more uniform across treatment modalities. Local tumor control and overall survival are the primary endpoints of treatment efficacy and the target sample size for such trials will be estimated so that the trial has sufficient statistical power to resolve what is judged to be a clinically relevant improvement in these parameters. From a normal tissue effects perspective, there are 2 major concerns. First, morbidity, especially late, is frequently inadequately recorded and reported. The lack of standardized scoring systems and the often poor quality control of morbidity scores limits the value of much of the published literature. Second, even when morbidity is reported, the statistical power of the trial is not sufficient to resolve a clinically important change in this endpoint. Thus, in most trials, the possible change in therapeutic gain cannot be reliably judged. This is clearly a field in which more research is urgently needed.

**Surrogate Markers and Endpoints for Late Effects**

The long latent period and the problems with low statistical power in many trials with late effects as a primary endpoint have stimulated interest in surrogate markers for late effects. It is useful to distinguish between surrogate markers and surrogate endpoints for late effects. A surrogate marker is a biological effect of treatment that, when it occurs in an individual patient, changes the probability that this patient develops a sub-
sequent clinically relevant late effect. The ideal surrogate marker would have both a high positive predictive value and a high negative predictive value for the clinical endpoint of interest. Also, detection of the surrogate marker should offer a clinically relevant lead time allowing a rational selection of cases for modification of ongoing primary therapy or for early interventions to relieve or prevent subsequent late morbidity.

One example of a surrogate marker is plasma transforming growth factor-β (TGF-β) as a marker for pulmonary injury after radiotherapy. A normalization of the plasma level of this marker toward the end of a course of radiotherapy was shown to yield a positive predictive value of 90% for identifying patients who did not develop radiation pneumonitis. In a series of 38 patients with inoperable non–small-cell lung cancer, Anscher et al used changes in TGF-β level as a criterion for escalating radiation dose in the individual cases. Twenty-four patients had persistently abnormal TGF-β levels and received a total dose of 73.6 Gy. Among 14 patients whose TGF-β levels were normal after 73.6 Gy, 8 were escalated to 80 Gy and 6 were escalated to 86.4 Gy. In the 86.4-Gy group, dose-limiting toxicity was reached because there were 2 (33%) grade 3 late toxicities. The authors concluded, from this uncontrolled study, that it may be feasible to use plasma TGF-β levels to select patients for radiation dose escalation.

Another example is from a prospective longitudinal study of clinical, endoscopic, and histopathologic rectal toxicity in 33 patients during ongoing pelvic radiation therapy. Although clinical symptoms of early rectal morbidity progressed toward the end of the 6-week treatment course, endoscopic pathology was maximal at 2 weeks and stabilized thereafter. Also, histologic changes were consistently more pronounced at 2 weeks than at 6 weeks. These observations could point to possible surrogate markers of late effects and may have implications for the design and timing of prophylactic and therapeutic interventions to reduce radiation proctitis.

The long latent period of late effects must also be taken into consideration when estimating the positive and negative predictive value of a surrogate marker. A patient expressing the marker who dies before reaching the clinical endpoint is not necessarily a false-positive; again, it is necessary to use actuarial methods to correct for censoring.

A surrogate endpoint, on the other hand, does not necessarily affect the probability of an individual expressing the clinical endpoint of concern but is indicative of treatment toxicity in a population of patients. For such an endpoint to be useful, it should provide a considerable lead time, should offer advantages in terms of reliable quantification and/or improved statistical power, and should be responsive to modifications of therapy. As an example, the incidence of confluent mucositis after radiation therapy in the head and neck region has been found to be a good indication of the overall biological intensity with respect to early morbidity. It is not a good surrogate endpoint for late effects as illustrated for example by the significant increase in the prevalence of confluent mucositis, but significant decrease of the incidence of late effects, after continuous hyperfractionated accelerated radiotherapy as compared with conventional radiotherapy for squamous cell carcinomas of the head and neck region.

Mechanistic Studies

Biological research on normal tissue effects ultimately aims to prevent or ameliorate the side effects of cancer therapy. Over the past decade, it has become clear that both early and late effects in normal tissues are a result of a complicated network of signaling molecules (cytokines) and receptors initiated at the time of radiation injury and progressing until the clinical endpoints are manifest. This recognition led to a rapid increase in biological studies attempting to identify the primary effectors in late-effect induction pathways that may then serve as targets for ameliorative treatments. However, the simple overexpression (or activation) of a moiety is not necessarily indicative of a causal relationship or even its role in the pathogenesis of a late effect. Indeed, the multifunctional nature of many cytokines, chemokines, and so on means that some may serve as both good guys and bad guys, being involved in both the induction of late effects, but also serving as necessary components of healing and repair in the normal tissue. A primary example of such a molecule is the previously mentioned TGF-β. In addition, many cytokines...
exist in a state of homeostasis, with the overexpression of one being balanced by the expression of another; such compensation mechanisms have been frequently observed and discussed among members of the interleukin family.39 Finally, consideration must be given to the observation that no 2 tissues have identical expression of late effects. This differentiation probably occurs as a result of variations in cell populations in each tissue, leading to different target cells. Thus, no assessment of the potential usefulness of a signaling factor can be made without some recognition of the spatial component (ie, the cells and tissue environment that are involved).

All of these considerations pose a great challenge to preclinical research in model systems. Obviously, results from in vitro studies in permanent cell lines must be transferred to primary cell lines or tissue cultures. In any case, conclusions from these studies have to be validated in vivo in experimental animals, using relevant endpoints and well-designed experimental protocols. The same applies to results from molecular biology studies. In many cases, in vivo experiments reveal that intervention in 1 pathogenic pathway, which has been shown to modulate the cellular response in vitro, is ineffective in vivo because of either alternative pathways or because of interactions between different cell populations present in a complete tissue.

Predictive Assays

The large patient-to-patient variability in the response to a course of radiotherapy has stimulated interest in predictive assays of both tumour and normal tissue radioresponsiveness. Again, it is useful to distinguish between surrogate markers for effect and predictive assays. Surrogate markers in vivo could in principle form the basis for a predictive assay. Such an approach would require that at least part of the therapy be delivered to the patient, which may limit its value in clinical decision making. Likewise, although the outcome of a predictive assay would affect our knowledge of the probability of a patient reaching a specific endpoint, it is not a surrogate marker of biological effect as discussed earlier.

Although several assays have been evaluated and at least some of these have shown promise, there is still no predictive assay that is anywhere near implementation in the clinical routine. In the absence of direct evidence for the utility of normal tissue predictive assays, it is of interest to estimate the relative importance of deterministic versus stochastic effects in explaining patient-to-patient variability in normal tissue responsiveness. A recent article40 analyzed the occurrence of telangiectasia of the skin in patients treated with bilateral internal mammary fields and found that as much as 90% (with 95% confidence limits 65% and 100%) of the variability in the radioresponsiveness in the right-sided field was explained by the radioresponsiveness in the left-sided field. This suggests that deterministic factors may dominate and stimulates further research to define these determinants.

Combined Modality or Intensified Radiotherapy

Novel combinations of therapies and new dose-volume combinations, radically different from those giving rise to most of our historical normal tissue complication data, may considerably change the incidence and type of morbidities seen. Interestingly, in the present context, this may lead to a need for changing the toxicity criteria themselves.

A concrete example is the Fox-Chase modification of the SOMA/LENT scale for rectal toxicity.41 This modification was suggested based on long-term follow-up of a group of patients with prostate cancer included in a nonrandomized dose-escalation study. The investigators observed a number of cases of chronic rectal bleeding requiring at least 1 blood transfusion and/or more than 2 coagulations after high-dose conformal radiotherapy for prostate cancer and proposed to classify this as grade 3 rectal toxicity. This manifestation of rectal damage was not included in the original SOMA/LENT system, simply because it is very rarely seen with conventional dose levels. Similarly, when 3-dimensional conformal radiotherapy is used to escalate the dose to small-volume lung cancers, it is likely that the dose-limiting toxicity will be structural damage to the lung and major vessels, rather than the restricted lung-function that limits the dose of conventional radiotherapy.

More recently, multimodality therapies with novel agents, including biologically targeted ther-
apies, are currently undergoing early clinical testing. There is little doubt that these strategies will improve tumor control and consequently also improve survival rates. However, this may come at a price: many of these new treatment regimens do involve an implicit or explicit acceptance of an increased incidence of treatment-related morbidity. Also, it can be expected that the morbidity profile may change as dose distributions are radically changed or drugs with new mechanisms of action are introduced.

**Conclusion**

An increasing number of new therapies and modifications of established therapies originate from advances in basic and clinical cancer research. The translation of these into clinical practice is a huge challenge also to the systematic scientific study of normal tissue effects. A therapeutic gain cannot be achieved without carefully balancing tumor cure and survival rates against morbidity and quality of life. Development of standardized common toxicity criteria and a widespread adoption of these in clinical trials would be a major step forward for clinical cancer research. For late effects, we suggest that the actuarial incidence is reported as a minimum and that this is supplemented by prevalence or cumulative incidence estimates if this is regarded relevant. Subjective symptoms, objective signs, management of toxicity, and analytical measures are complementary dimensions of toxicity, and we propose that these are reported and analyzed separately. In mechanistic studies or studies testing interventions to prevent or ameliorate morbidity specific biological endpoints are preferable. An involvement of normal tissue radiation biologists should be sought to optimize the design of clinical instruments for treatment effect assessment. A consensus is needed for the development of guidelines for the analysis and reporting of normal tissue effects.

**References**

7. Pav JI, Denekamp J, Letchert J: Late effects damage scoring: The SOMA scale. Radiother Oncol 35:11-60, 1995
20. Bentzen SM, Turesson I, Thames HD: Fractionation sensitivity and latency of telangiectasia after postmastec-


