Review Article

Surrogate endpoints in early prostate cancer research

Scott Williams

Division of Radiation Oncology and Cancer Imaging, Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, Australia

Correspondence to: Dr. Scott Williams, MBBS, MD, FRANZCR. Division of Radiation Oncology and Cancer Imaging, Peter MacCallum Cancer Centre, 305 Grattan St., Victoria 3000, Australia. Email: Scott.Williams@petermac.org.

Abstract: Clinical research into clinically-localized prostate cancer (PC) is a highly challenging environment. The protracted durations and large numbers required to achieve survival endpoints have placed much pressure on validating early surrogate endpoints. Further confounding is the predominance of deaths from causes other than PC. The analysis of multiple randomized clinical trials in early PC has shown MFS to be a robust surrogate for OS, using a contemporary analytic framework that identify patient-level and trial-level associations. This could potentially save around one year of trial follow-up in some therapies. Identification of a similarly robust surrogate at a substantially earlier timepoint remains a major challenge. Multiple biochemical indices based on PSA have been proposed in the literature, but all remain to be validated at the trial-level. Operationally, many of these indices have inherent biases such as immortal-time bias (ITB) and interval censoring that potentially weakens associations and the individual- or trial-level. The complexity of a failure definition can also impact the reliability of the derived outcomes. Confounding issues such as the impact of comorbidities leading to non-cancer deaths have been largely dealt with by their exclusion using cancer-specific endpoints and advanced statistical methods, while issues such as PSA "bounce" and recovery from androgen deprivation therapy remain important to account for in cohorts treated with radiotherapy. Several potential surrogate endpoints based on serum prostate-specific antigen (PSA) levels show promising associations with PC-specific and overall survival (OS) in individual studies. Further large collaborative projects will continue to refine potential indices with these issues in mind, and explore the objective of an early surrogate of OS.

Keywords: Prostate cancer (PC); surrogate markers; biochemical failure

Submitted Jan 22, 2018. Accepted for publication May 08, 2018.
doi: 10.21037/tau.2018.05.10
View this article at: http://dx.doi.org/10.21037/tau.2018.05.10

Introduction

Clinical research involving the early phases of prostate cancer (PC) is embattled by the typical long natural history of the disease. Survival endpoints, although valid, reliable and meaningful, routinely require large and protracted studies to have adequate statistical power. For example, studies of greater than 1,500 men followed for over 10 years being routinely required to confidently identify (or refute) a survival impact from a new therapeutic approach to newly diagnosed PC (1,2). Even in the recurrent disease setting with established biochemical failure (bF) based on prostate-specific antigen (PSA) dynamics, large protracted studies are required to evaluate survival endpoints (3). This places significant delays in developing the evidence base to inform practice, as well as imposing a substantial logistic and financial burden on research organizations. Hence, the identification of surrogate endpoints that offer efficiencies in terms of smaller or shorter trials is paramount.

Inevitable with the long history of most PC is the impact of death from other causes in this predominately older population. It is well described that comorbid conditions account for the vast majority of deaths in those diagnosed with PC at age 70 or greater. Even in younger men, comorbidity contributes considerably to overall survival (OS) regardless of disease biology (4,5). This then has the effect of diluting the ability to observe a true treatment-related outcome effect, weakening statistical power and making event rate predictions less reliable when planning studies. Improvements in the therapy of non-cancer conditions over time can impact many assumptions about event rates that are as fundamental to a studies success as improvements to oncological outcomes over time.

In this review, the current contemporary framework to define surrogacy will be briefly outlined along with the historical context from which it developed. Performance issues of purported surrogates will be discussed, particularly with respect to the potential biases and operational issues inherent in various approaches. Largely, published work will be categorized into classes of design, along with summaries of the performance characteristics and applicability to the surrogacy framework. Specific studies will largely be used to illustrate issues or successes of note only. Overall, the focus will be on data pertaining to early PC, from the time of diagnosis to the identification of various PSA-based criteria, with less emphasis on surrogates that utilize features of clinical progression (such as metastasis). Furthermore, as stated in several seminal biochemical failure definition publications (6,7), the objective of these early indices is to find a population average measure of outcome that is associated with a later, more clinically relevant, outcome which also effectively captures treatment effect. These indices are not designed to be, and should not be used as, tools with which to make clinical decision. The identification of recurrence should continue to be based on the individual's clinical features available to the physician.

Defining surrogacy

A surrogate endpoint is one which is typically not inherently meaningful in itself, such as biochemical progression, but can provide reliable evidence that a treatment will have an impact of "hard" outcomes such as how patients feel, function or survive. It should have a scientifically plausible link to the true outcome, and ideally, the surrogate should be embedded within the causal pathway of the disease progression to the true endpoint (8). The use of surrogate endpoints is becoming more commonplace; in the 5 years to 2014, two-thirds of the 55 oncology agents approved by the US Food and Drug Administration (FDA) were on the basis of surrogate endpoints (response or progression-based), including all 25 of the accelerated approvals (9).

The framework with which to evaluate indices as potential surrogate endpoints has evolved over recent decades, with the seminal work defining the field by Prentice in 1989 (10). That work proposed that four simultaneous criteria need to be met to support surrogacy: (I) that treatment has a significant impact on the surrogate endpoint; (II) that treatment has a significant impact on the true endpoint; (III) that the surrogate and true endpoints are correlated; (IV) that the full effect of treatment on the true endpoint is captured by the surrogate. This final criterion has been argued as unduly restrictive, with very few indices able to comply.

Building on this, Buyse *et al.* (11) proposed a meta-analytic method of surrogacy evaluation where both a strong association of surrogate to true endpoint is seen across individuals, as well as the treatment effect being correlated across studies. Association of the surrogate with the true endpoint in the individual only requires data of one study, while the treatment effect estimation involves assessing the effect size on the surrogate against that seen on the true endpoint for multiple randomized controlled trials RCTs. A linear relationship enables the determination of a "surrogate threshold effect" (STE), which is the smallest treatment effect on the surrogate that will confidently give a non-zero effect on the true endpoint (12). This depends on the strength (variance) of the association of treatment effects between surrogate and true endpoints, while the coefficient (slope) of the weighted linear regression enables a degree of prediction of the effect size on the true endpoint from the surrogate effect size.

This form of rigorous testing across multiple RCTs is becoming more widespread (13-16). However, of the 55 FDA approvals noted earlier, 65% were not supported by meta-analytical validation data of the surrogate endpoint (9). Mandatory monitoring following FDA drug approval initially based on these unvalidated surrogate endpoints has now led to withdrawal of approval of at least one drug (17), reinforcing that the simplistic identification of association does not equal surrogacy.

Surrogates of OS

The most comprehensive effort to evaluate surrogate endpoints is PC is the Intermediate Clinical Endpoint in Carcinoma of the Prostate (ICECaP) initiative. To address the requirements of a meta-analytic approach to surrogate evaluation, a systematic review identified 102 potentially eligible randomized trials globally that ultimately led to the data of 28,905 patients from 43 studies being contributed and available for individual patient and trial-level analyses. Initial analyses focused on clinical progression surrogates of OS, namely disease-free survival (DFS; loco-regional and/or distant metastatic relapse) and metastasis-free survival (MFS; distant metastases only). For MFS, the analysis was limited to 12,712 patients from 19 studies that collected metastasis data. The majority of evaluable randomised studies were of high-risk primary disease treated with RT.

As described, the aim was to fulfil two conditions of surrogacy. The first, that of correlation of the intermediate clinical endpoint (ICE) and OS in the individual, was strong between either DFS or MFS and OS (kendall's tau =0.85 and 0.91 respectively). Similarly, the R^2 between 5-year DFS or MFS and 8-year OS was high at 0.86 and 0.83 respectively (18). The second condition required the demonstration of correlation between the treatment effect on both the ICE and OS. Weighted linear regression of the hazard ratio (HR) found for the ICE against the HR found for OS showed a strong correlation between MFS and OS (R^2=0.92), which substantially reduced (R^2=0.73) for comparison of treatment effect between DFS and OS. The estimated STE was 0.88 for MFS, suggesting that any treatment effect (HR) on MFS larger than this would predict a non-zero effect on OS (18). Surrogacy was maintained across a variety of primary and adjuvant therapies. DFS was consistently weaker as a surrogate of OS, possibly due to more indolent local failures being added into the failure definition.

Biochemical indices as surrogate endpoints

PSA-based endpoints have been widely adopted as early outcome indices in radical therapy approaches such as radiation therapy and prostatectomy, based on associations with subsequent clinical events in large retrospective databases (6,7). Extensive work has gone into the development of similar indices for use in the CRPC space (19,20). Unfortunately, no biochemical index has been validated as a surrogate of clinical failure adequately enough to be approved by regulatory bodies (21). There, however, remains a strong desire to utilize PSA indices as early surrogates of survival events due to their potential to identify treatment effects relatively rapidly. In practice, several categories of PSA indices can be described ranging from simple measures of absolute PSA level through to complex dynamic or longitudinal indices.

Absolute PSA

The use of an absolute PSA level to identify disease progression after radical therapy has been extensively tested, particularly in the post-prostatectomy setting, with the most influential data being that of Stephenson *et al.* (7). They analysed the data of 3,125 men treated surgically and followed for a median of 49 months. The "true" endpoint was metastatic progression (MP), which had occurred in 75 men (2.4%). They described 6 variations of absolute PSA level: a single PSA level of ≥0.2, 0.4, or 0.6 ng/mL; and a PSA ≥0.1, 0.2, or 0.4 confirmed with a subsequent rise. The biochemical failure definition (bFd) was handled as a time-dependent covariate in a Cox model with MP as the outcome, and the R^2 goodness-of-fit used to rank the different bFd's. They concluded that the optimal early predictor of subsequent clinical progression in this cohort was a PSA ≥0.4 and rising, that is, confirmed by another higher level at a later time. While this bFd achieved the highest R^2, this was modest at 0.21 (on a range of 0–1, with 1 being perfect accuracy). Although the definition involves a PSA at or above 0.4 ng/mL, the median PSA level at bF was informative: for the single PSA ≥0.4 ng/mL bFd this was 0.57 ng/mL, while the addition of a confirmatory rise increased the median PSA at bF to 1.0 ng/mL. The median time to bF was extended by 6 months also in waiting for the confirmatory rise. More recently however, a large mature cohort (n=13,512, median follow-up =9.1 years) has suggested that a level of 0.4 ng/mL without a subsequent confirmatory rise is optimal (22).

Absolute PSA levels have historically not been recommended as bFd's after EBRT however. This is largely due to EBRT having nadir PSA (nPSA) levels that were often in the range of 0.2–0.4 ng/mL, hence making absolute PSA level bFd's prone to poor accuracy around these levels. Indeed, a PSA ≥0.2 had a sensitivity of just 0.09 in an EBRT cohort, rising only marginally to 0.26 at PSA ≥0.5 (23). For more ablative radiation approaches however, such as combined EBRT and brachytherapy, data have been put forward to support the use of a low absolute PSA (≤0.2 ng/mL) as a reliable index of long-term outcome (24).

Relative PSA rise

"Relative" bFd's utilize an absolute PSA level that is internally calibrated in some manner. Mostly, an absolute PSA rise relative to the nPSA level has been investigated. This has been of most interest in the RT research space, where it was suggested that a bFd that controlled for the inherent variation in PSA nadir levels across individuals may be more accurate. This has been extensively examined in the RT and also RP cohorts. In an analysis of 4,839 men with a median follow-up of 6.3 years, Thames *et al.* (23) showed optimal sensitivity and specificity related to clinical failure were obtained with a PSA ≥2 or 3 ng/mL above the lowest PSA to date. Understandably, increasing the PSA level above nadir increased specificity (nadir +5 ng/mL bFd specificity =0.94), but lowered sensitivity. Ultimately, the optimal accuracy was obtained with nadir +2 ng/mL, which was subsequently widely known as the "Phoenix" definition of bF and ratified as a research endpoint by RTOG-ASTRO (6). This definition has been associated with OS in retrospective data (25).

Testing a similar approach in the post-prostatectomy setting has been performed (26) and showed the nadir +2 ng/mL bFd (N2D) delayed the identification of bF by approximately 5 years beyond a definition of ≥0.2 ng/mL. This substantial delay has been argued as unjustified and also prone to bias due to many men undergoing salvage therapy prior to a PSA of nadir +2 ng/mL (7).

PSA nadir

The nPSA level has been suggested to have a significant impact on outcome in both the surgical (27,28) and radiation therapy (29,30) groups in multiple studies. Nadir levels are used as continuous covariates in models, or dichotomized as detectable or undetectable based on a variety of levels. Strong association with clinical outcomes is often shown in the papers and the nPSA is often proposed as a predictor of need for salvage therapy (31). Two publications have now detailed PSA nadir for surrogacy using randomized trial data (30,32). Combining the data of two large trials comparing RT alone to RT + 6 months of ADT, the achievement of a PSA nadir of 0.5 ng/mL was analysed in relation to prostate cancer-specific mortality (PCSM) using a time-dependent competing risks model. PSA nadir was strongly associated with PCSM, and the proportion of treatment effect explained by the surrogate was >80%, with less than 2% of the treatment effect persisting after inclusion of the surrogate, satisfying Prentice criteria (30,32). Using a similar approach, analysis of a subset of 157 men enrolled in a randomized trial of RT +/– ADT, a PSA nadir >0.5 ng/mL was associated with worse OS (30) and explained enough of the treatment effect to satisfy Prentice criteria also (32), suggesting utility as an early surrogate of PCSM and perhaps OS.

Time to nPSA is often suggested to be a valid surrogate of outcome (33-35), but is frequently compounded by bias (36) (discussed later). By conducting a landmark analysis to minimize such bias, Skove *et al.* (28) found that a time to nadir inside 3 months was associated with subsequent bF less often than an nPSA between 3 and 6 months in a retrospective post-prostatectomy dataset.

Dynamic PSA algorithms

Dynamic PSA algorithms range from the relatively simples, such as the three PSA rise bFd used for RT outcomes (6), to those that utilize very complex longitudinal models (37). Algorithms that define failure within a certain time period also fall into this category, as they indirectly specify a rate of change.

The original ASTRO bFd was a simplistic dynamic algorithm, simply specifying three consecutive PSA rises irrespective of rise amount and timing, based on the high likelihood of further rises once this occurred (38). Backdating was included in this original definition, introducing many statistical concerns (discussed subsequently). The most complex dynamic algorithms to date are based on linear mixed modelling of the PSA over time in those treated with RT and no additional ADT. Incorporating initial PSA, clinical stage, Gleason score and radiation dose as covariates, the model can accurately predict impending PSA behavior (39,40). This model has been extended to jointly incorporate clinical failure risk based on these predictions, enabling the risk of clinical failure in future years to be predicted using an online calculator (37,41,42). While potentially very useful clinically in the individual, such models would be difficult to implement as a more general surrogate endpoint.

The rate of PSA rise, typically expressed as a PSA doubling time (PSAdt), has been extensively examined, including using data of several RCT's. In RTOG 92-02, PSAdt was calculated using PSA data to the time of bF (3 rise ASTRO Consensus definition) and the four Prentice criteria evaluated against cancer-specific survival (CSS). Criterion 1–3 were satisfied, however criterion four, that of the surrogate fully explaining the effect of treatment on CSS, was unable to be met (43). In a detailed analysis of the data of an Australian RCT, a PSA doubling time of <12 or <15 months (calculated with PSA data to the time of initiating salvage therapy or censoring) met all Prentice criteria as a surrogate for PCSM (44). These data suggest that PSAdt early in disease progression can be difficult to relate to cancer-specific death, while dynamics assessed at the time of salvage therapy initiation are more informative.

Surrogacy for OS was not assessed in these studies.

While the actuarial analysis of bF best enables a hazard ratio to be derived in a RCT, there has been much impetus to look for a more dichotomous short-term endpoint to enable rapid translation of research questions. Several investigators have analysed the presence of bF at various time points; in essence, these are dynamic indices that integrate a certain amount of PSA rise within a certain timeframe. Buyyounouski *et al.* showed with that time to bF (TTbF) using the nadir + 2 ng/mL definition of <18 months is associated with a higher risk of metastases and death from cancer in separate discovery and validation cohorts (45,46). Denham *et al.* further validated this with RCT data, showing that a TTbF of <18, <24 or <30 months all met Prentice criteria for surrogacy of CSS in this trial of RT +/– ADT (44).

Performance issues with biochemical indices

In discussing the performance of PSA-based indices as possible surrogates, the potential for bias within the result must be examined critically. Several categories of these have been well described.

Immortal time bias

ITB is a common and insidious issue throughout much of oncology, including within many analyses of PSA data. Its presence in the scientific literature has been noted as early as 1844, where it was shown that Catholic Popes lived significantly longer than their artist contemporaries. At the time this was promoted as unquestionable support for a virtuous lifestyle, without accounting for the bias that one had to live achieve a certain degree of longevity prior to being ordained—which has been, with more appropriate analysis, suggested as erroneous (47).

Within PSA data, ITB can present in many ways. Possibly the easiest to identify is when "time to…" indices are proposed, such as time to nPSA. Analyses have shown prolonged time to nadir to be strongly associated with improved subsequent outcomes (33,48,49); all of which included the time to nadir as a covariate in a fixed proportional hazards model, that is, it is included along with other fixed pre-treatment covariates used to calculate outcome from time of therapy. Clearly, bias plays a role here as, by definition, bF or other clinical failure endpoints cannot have occurred prior to nadir. So those with a longer time to nadir will have an inbuilt trend to later failure. Similarly, throughout the duration to nadir occurring, the patient must also have been alive, and hence is technically immortal until nadir was reached. This has then led to data showing that the majority of the effect of time to nPSA, when included as a baseline covariate, is due to ITB (36). At a minimum, more emphasis should be given to studies that utilized landmark analysis or the incorporation of time to nadir as a time-dependent covariate to try to understand this phenomenon (50). Others have developed more complex longitudinal models of PSA behavior that can be jointly modelled with clinical failure to avoid this issue (41).

Not immediately recognized was the strong ITB inbuilt in the original ASTRO Consensus definition of bF. Backdating the date of failure to the midpoint between nadir and first PSA rise following identification of three PSA rises creates a strong bias. The patient is hence not only alive at the time of failure, but also up to the time of the third PSA rise—hence imparting an immortal period after bF. This creates a disconnect between the hazard of clinical failure (steadily present from time zero) and that of bF (zero to the time of the third PSA rise), which is difficult to correct statistically if an association between biochemical and clinical failure is being sought (25). Other prominent examples exist, such as the association of duration of ADT received and clinical failure (51). It should also be noted that any indices using the nadir level or date must also invoke a small degree of ITB, as to identify the nadir point (assuming the patient does not have an undetectable PSA level), it can only be done in retrospect. This is because a subsequent rise, or several rises, must have occurred to then be confident that any low PSA level is the true nadir. The duration between the nadir and the subsequent time at which it is confidently identified is immortal and failure-free.

Detection biases

Inherent in any failure definition, whether biochemical or clinical, is a characteristic interaction with detection biases. These biases distil down to the simple phenomenon of "the more often you look, the more often you find". These biases impact strongly impact biochemical outcomes in particular via PSA test frequency and regularity (or adherence to protocol testing), along with follow-up duration adequacy.

These issues have been examined in detail in a series of analyses coinciding with the development of the Phoenix definition of bF in the post-radiotherapy setting (52,53). Several bF definitions representing various categories of bFd of interest were examined: the existing "ASTRO Consensus" definition (ACD) defined as 3 consecutive PSA rises with failure backdated to the midpoint between nadir and first PSA rise (38); the "Three rise" definition (3RiseD) analogous to the ASTRO definition but without backdating; the "Phoenix" definition defined as a PSA rise of 2 ng/mL above the nadir (N+2D) (6); and a simple definition of an absolute PSA of 3 ng/mL (Abs3D) (52).

Each of these bFds were examined for dependence on various aspects of PSA testing. For PSA test frequency reliance, a simple comparison of the crude mean time between testing (MTBT) showed that those having infrequent PSA testing (MTBT >8.2 months) had a hazard ratio (HR) of bF that was 67–78% less than those with a MTBT <5 months when examined across all definitions. To control for selection bias in these data, where more aggressive cancer may have had more intense follow-up, a subset of men that had very regular frequent follow-up (MTBT =3.3 months) were examined with all data or with every second test removed (creating a MTBT =6.7 months). In this small subset (n=107), the median time to bF was extended by approximately 3 months in the N+2D and Abs3D, however, the 3RiseD was pushed out by 8.2 months. Backdating in the ACD brought this back to 5.1 months difference, strongly suggesting that test frequency is an important feature in the calculation of bF outcomes via any class of definition (52).

Another concern for PSA testing in particular is the impact that irregular testing—either additional or missing tests—has on outcomes. To assess this, Williams (53) developed a simple error term [the Irregularity Index (II)] to denote how concordant a patients PSA testing history was with perfectly regular testing. The II derived a HR of 0.40–0.47 across all definitions when patients in the quartile that had the most regular testing were compared with those with the least regular testing quartiles, with the predominant effect seen in the most irregular testing. Overall, both PSA testing frequency and regularity were as influential on the derived bF outcome as conventional prognostic factors, suggesting that indices of PSA testing need to be described if outcomes are to be reliably compared across studies.

One final and similarly important issue that can be grouped with detection biases is that of follow-up duration adequacy. In deriving an actuarial failure estimate, the most important issue in arriving at an accurate estimate of failure hazard is the relative timing of censoring and failure events. Having a follow-up duration that allows the majority of censoring events to fall beyond the median time to failure is desirable. By analyzing outcomes based at annual follow-up intervals that varied median follow-up between 36 and 96 months, it was shown that non-backdated bFds behaved in a consistent and reliable manner (52). Apart from the earliest time points (amounting to a median follow-up <60 months), no significant differences in outcome estimates were discernable. The estimates at earlier times were conservative, modestly over-estimating failure when extrapolating to time points well beyond median follow-up. The backdated ACD, however, performed poorly by comparison, as shown in earlier studies (54-57). Early analysis points were overly optimistic in their estimates, with the freedom from bF falling incrementally with each additional year of follow-up. This reduced the initial 5-year freedom from bF (FFbF) estimate of 55% (at a median follow-up of 36 months) down to 37% at a median follow-up of 96 months. Importantly, a simple relationship between crude time to failure (TTF) and time to censoring TTC) could be used to determine adequate follow-up reliably in the prospective bFds (median TTF <25th centile of TTC; or median TTC > median TTF ×1.6). No reliable rules for follow-up adequacy in the ACD could be found however, leading to subsequent recommendations that ACD outcomes only be reported at time points at least 2 years short of the median follow-up (6).

Impact of false positive issues

Particularly in the case of patients treated with RT approaches, several benign phenomena of PSA irregularity following treatment are well described. Following low-dose-rate brachytherapy (LDRBT) in particular, a non-sustained PSA rise will be seen in approximately 30–50% of cases, typically 12–18 months post-therapy (58-62). This PSA "bounce" usually lasts less than a year and is associated with a good prognosis, although strongly contaminated by ITB in many studies. Similarly, a prominent PSA rise can be seen on recovery of testosterone following ADT given in combination with RT. From a highly suppressed PSA (often undetectable) while on ADT, there is often a series of PSA rises prior to a PSA plateau following testosterone recovery. This results in high rates of false calling of biochemical failure with a 3 rise bFd (24% in one study) which is much reduced with a N+2D (2%) (58). While highly dependent on the bFd used, issues that alter PSA significantly without altering the risk of failure clinically will weaken the ability to define a biochemical surrogate of clinical failure.

Human errors of interpretation

Validation of laboratory biomarkers routinely involves extensive testing to determine parameters such as reliability, reproducibility and calibration across all possible usage scenarios. This is rarely, if ever, done with bF indices despite our clear desire to use them as a biomarker.

An important feature of any bFd is reliability in interpretation when implemented by what may be a large number of users of varying experience and capacity to perform detailed quality assurance of their data. Complex algorithms or those which rely on manual curation of data have poor inter-observer performance. In one illustrative study, identical data of 1,200 men treated with a variety of radiation therapy techniques were analysed at four separate institutions experienced in publishing PC outcomes. Each institution assessed the data for bF using ASTRO definition, Phoenix definition and absolute PSA ≥3 ng/mL endpoints. The backdated ASTRO definition had poor consistency, with 5 year freedom from bF (FFbF) ranging from 49.8% to 60.9% for the cohort depending on the institution performing the calculation (63). Only 87% of cases had a consistent bF status calculated across institutions using the ASTRO definition, while the other crucial factor for actuarial analyses—time to failure or censoring—showed substantial variation between institutions (≥2 months) in 23% of men.

The other two prospective bFds had >92% concordance in bF identification, but intriguingly, still had major variation in the failure call time in 21–36% of bF cases (63). Issues such as how PSA "bounce" data was handled were highly influential. Thus, despite being seemingly simple failure definitions, the 15 variations of interpretation of the definitions noted in that paper plus earlier work (53), make it clear how poorly PSA endpoints can behave due to human interpretation issues. As illustrated by that work (63), scrutiny of performance and publishing of analytic algorithms accounting for common areas of misinterpretation are crucial to good bFd performance.

Discussion

With the long natural history of PC, many attempts have been made to define a clear surrogate of clinical outcomes that can serve to accelerate the adoption of new therapies. Although numerous indices have been suggested as surrogates, few studies to date have validated contemporary criteria required for a surrogate marker to be recognized by regulatory bodies. Significant inroads have been made with surrogates based on early post-therapy response or failure endpoints, but only in relation to outcomes removed of comorbid disease influence by either the use of a cancer-specific endpoint or a competing risks model (or both) (32,44). It is the impact of comorbid illness that represents the greatest difficulty, as comorbid deaths are likely to contribute the majority of deaths seen in any cohort, even in aggressive disease (4). It is understandable that the removal of this confounder has been the objective of many researchers—assuming that if a patient has zero hazard of death from other causes (i.e., would otherwise live forever), then identification of cancer recurrence at any time would be ultimately always be associated with a survival detriment and surrogacy would be identifiable. Removing the effect of comorbidity can then better allow associations of early response or failure to be identified which is useful to confirm biological hypotheses, but does not inform us well about the potential impact of a new therapy on OS.

For the optimal endpoint of OS, only the data related to MFS from the ICECaP initiative have shown robust patient and trial-level surrogacy (18). The identification of this surrogacy required the collective analysis of many thousands of individual patient's data across multiple clinical trials with prolonged follow-up. Sobering is the observation that even when a reasonably strong treatment effect is anticipated (a hazard ratio OS =0.67), the likely time saving in using a MFS endpoint would only be approximately 1 year (18). Comorbid issues continue to impact even at these late time points—MFS is not a perfect surrogate of OS, meaning other causes of death are also at play. Also confusing this are the "moving goalposts" with improvements in management of progressive disease, with a multitude of life-prolonging options influencing the time between progression and death (64-67).

Key principles surrounding surrogate marker performance still require very careful evaluation before a surrogate can be accepted as robust. As summarised previously (68), these are (I) analytic—to ensure validation of the process to determine the intermediate endpoint result, (II) statistical—to ensure acceptable patient and trial-level associations, and (III) utilization—to determine whether the surrogate has context-specific limitations.

Analytic validation is a major challenge for any bFd proposed as a surrogate as, previously, much precious statistical power lost due to biases inherent in definitions to date. As described, variations in the frequency or regularity of PSA testing can strongly influence the result in some bFds (52). Although these detection biases (interval censoring) are minimized when comparing randomly allocated arms on protocol, they may hamper interstudy comparisons and thus overall reliability. The most troublesome is that of immortal time bias, as it is often underappreciated in both presence and impact. Essentially, all retrospective bFds contain some degree of ITB that will weaken the statistical model of association with the true endpoint. Conversely, prospective or "real-time" bFds are those that can identify failure without requiring any additional time to pass—time during which the patient cannot suffer an event by definition. Misinterpretation of bFds at a human level are analogous to inter-laboratory variations, and can have a significant impact on the calculated surrogate outcome. Simple bFds are less error-prone and hence desirable from this perspective (52).

Statistical validation of surrogacy in the contemporaneous setting is predominately based on the confirmation of individual- and trial-level associations as described by Buyse *et al.* (11). The association of the surrogate with patient outcome can be attained with data from just a single clinical dataset or study. Using appropriate statistical methods, many possible surrogate indices have been shown to fulfil this criterion ranging from early response markers to failure definitions (25,32,46). Do date, no early surrogate has satisfied trial-level association criteria with OS in a manner analogous to the recently shown MFS surrogacy (18) however. The biases previously discussed combine with the impact of non-cancer deaths to dilute the potential to see an association, although no appropriately large-scale trial-level association has been performed to date. The derivation of a per-trial hazard ratio is ideal in this analytic framework, and thus surrogates that incorporate all available actuarial data will likely prove maximally powerful (69).

Utility of a surrogate is also a complex challenge for most tumour types, and PC is no exception. There are clearly many issues that need to be resolved to enable a robust early surrogate of OS to enter clinical trial design. Centrally, almost two decades ago now the International Conference on Harmonisation (ICH) suggested that, along with the surrogate being linked to both clinical outcomes and treatment effect, it should also demonstrate biological plausibility (70). Ideally, this requires the surrogate to be causally linked with the true endpoint whereby the surrogate lies directly within the progression pathway, and is also capture the mechanism of action of the treatment in question. Generalisability to other treatment scenarios cannot be directly assumed because of this, so treatments such as immunotherapy may not be well served by a surrogate endpoint validated for ADT. Although a good marker of PC activity after therapy in most cases, PSA does not directly cause cancer progression in itself and hence PSA-based surrogates will always start from a compromised position in this regard compared with clinical progression indices.

In the recent ICECaP analysis, MFS showed no evidence of interaction with the initial treatment modality (surgery or RT), suggesting MFS as a surrogate is generalizable to a broad array of clinical contexts presently (18). Being able to have harmonization of an early surrogate definition across treatment modalities would also be desirable, it would require careful validation. Prostatectomy removes the confounding factor of benign PSA activity that needs to be accounted for in post-RT bFds that are especially troublesome following ADT combined with RT or brachytherapy. Arriving at a consistent definition of early progression across such fundamental variations due to treatment modality remains a challenge.

Conclusions

The analysis of multiple RCTs in early PC treated with radical intent has MFS to be a robust surrogate for OS, and its use could potentially save around one year of trial follow-up in some cases. Identification of a similarly robust surrogate at a substantially earlier timepoint remains a major challenge. Multiple possible biochemical indices based on PSA have been proposed, but all remain to be validated, particularly at the trial-level. Confounding issues such as the immense impact of comorbidities leading to non-cancer deaths are yet to be adequately dealt with apart from their exclusion using cancer-specific endpoints and advanced statistical methods. Large collaborative analysis projects will continue to explore these issues in detail.

Acknowledgements

None.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Hamdy FC, Donovan JL, Lane JA, et al. 10-Year Outcomes after Monitoring, Surgery, or Radiotherapy for Localized Prostate Cancer. N Engl J Med 2016;375:1415-24.
2. Jones CU, Hunt D, McGowan DG, et al. Radiotherapy

and short-term androgen deprivation for localized prostate cancer. N Engl J Med 2011;365:107-18.

3. Shipley WU, Seiferheld W, Lukka HR, et al. Radiation with or without Antiandrogen Therapy in Recurrent Prostate Cancer. N Engl J Med 2017;376:417-28.
4. Albertsen PC, Hanley JA, Fine J. 20-year outcomes following conservative management of clinically localized prostate cancer. JAMA 2005;293:2095-101.
5. Ng SP, Duchesne G, Tai KH, et al. Support for the use of objective comorbidity indices in the assessment of noncancer death risk in prostate cancer patients. Prostate Int 2017;5:8-12.
6. Roach M 3rd, Hanks G, Thames H Jr, et al. Defining biochemical failure following radiotherapy with or without hormonal therapy in men with clinically localized prostate cancer: recommendations of the RTOG-ASTRO Phoenix Consensus Conference. Int J Radiat Oncol Biol Phys 2006;65:965-74.
7. Stephenson AJ, Kattan MW, Eastham JA, et al. Defining biochemical recurrence of prostate cancer after radical prostatectomy: a proposal for a standardized definition. J Clin Oncol 2006;24:3973-8.
8. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Stat Med 2012;31:2973-84.
9. Kim C, Prasad V. Strength of Validation for Surrogate End Points Used in the US Food and Drug Administration's Approval of Oncology Drugs. Mayo Clin Proc 2016;91:713-25.
10. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. Stat Med 1989;8:431-40.
11. Buyse M, Molenberghs G, Burzykowski T, et al. The validation of surrogate endpoints in meta-analyses of randomized experiments. Biostatistics 2000;1:49-67.
12. Buyse M, Molenberghs G, Paoletti X, et al. Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. Biom J 2016;58:104-32.
13. Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. J Clin Oncol 2007;25:5218-24.
14. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet 2014;384:164-72.
15. Imai H, Mori K, Wakuda K, et al. Progression-free survival, post-progression survival, and tumor response as surrogate markers for overall survival in patients with extensive small cell lung cancer. Ann Thorac Med 2015;10:61-6.
16. Zer A, Prince RM, Amir E, et al. Evolution of Randomized Trials in Advanced/Metastatic Soft Tissue Sarcoma: End Point Selection, Surrogacy, and Quality of Reporting. J Clin Oncol 2016;34:1469-75.
17. Carpenter D, Kesselheim AS, Joffe S. Reputation and precedent in the bevacizumab decision. N Engl J Med 2011;365:e3.
18. Xie W, Regan MM, Buyse M, et al. Metastasis-Free Survival Is a Strong Surrogate of Overall Survival in Localized Prostate Cancer. J Clin Oncol 2017;35:3097-104.
19. Scher HI, Halabi S, Tannock I, et al. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the Prostate Cancer Clinical Trials Working Group. J Clin Oncol 2008;26:1148-59.
20. Scher HI, Morris MJ, Stadler WM, et al. Trial Design and Objectives for Castration-Resistant Prostate Cancer: Updated Recommendations From the Prostate Cancer Clinical Trials Working Group 3. J Clin Oncol 2016;34:1402-18.
21. Gomella LG, Oliver Sartor A. The current role and limitations of surrogate endpoints in advanced prostate cancer. Urol Oncol 2014;32:28.e1-9.
22. Toussi A, Stewart-Merrill SB, Boorjian SA, et al. Standardizing the Definition of Biochemical Recurrence after Radical Prostatectomy-What Prostate Specific Antigen Cut Point Best Predicts a Durable Increase and Subsequent Systemic Progression? J Urol 2016;195:1754-9.
23. Thames H, Kuban D, Levy L, et al. Comparison of alternative biochemical failure definitions based on clinical outcome in 4839 prostate cancer patients treated by external beam radiotherapy between 1986 and 1995. Int J Radiat Oncol Biol Phys 2003;57:929-43.
24. Critz FA. A standard definition of disease freedom is needed for prostate cancer: undetectable prostate specific antigen compared with the American Society of Therapeutic Radiology and Oncology consensus definition. J Urol 2002;167:1310-3.
25. Williams SG, Duchesne GM, Millar JL, et al. Both pretreatment prostate-specific antigen level and posttreatment biochemical failure are independent predictors of overall survival after radiotherapy for prostate cancer. Int J Radiat Oncol Biol Phys 2004;60:1082-7.
26. Nielsen ME, Makarov DV, Humphreys E, et al. Is it possible to compare PSA recurrence-free survival after surgery and radiotherapy using revised ASTRO criterion--"nadir + 2"? Urology 2008;72:389-93; discussion 394-5.
27. Moreira DM, Presti JC Jr, Aronson WJ, et al. Natural

history of persistently elevated prostate specific antigen after radical prostatectomy: results from the SEARCH database. J Urol 2009;182:2250-5.

28. Skove SL, Howard LE, Aronson WJ, et al. Timing of Prostate-specific Antigen Nadir After Radical Prostatectomy and Risk of Biochemical Recurrence. Urology 2017;108:129-34.
29. Ray ME, Thames H, Levy L, et al. PSA nadir predicts biochemical and distant failures after external beam radiotherapy for prostate cancer: A multi-institutional analysis. Int J Radiat Oncol Biol Phys 2006;64:1140-50.
30. Royce TJ, Chen MH, Wu J, et al. Surrogate End Points for All-Cause Mortality in Men With Localized Unfavorable-Risk Prostate Cancer Treated With Radiation Therapy vs Radiation Therapy Plus Androgen Deprivation Therapy: A Secondary Analysis of a Randomized Clinical Trial. JAMA Oncol 2017;3:652-8.
31. Lamb DS, Denham JW, Joseph D, et al. A comparison of the prognostic value of early PSA test-based variables following external beam radiotherapy, with or without preceding androgen deprivation: analysis of data from the TROG 96.01 randomized trial. Int J Radiat Oncol Biol Phys 2011;79:385-91.
32. D'Amico AV, Chen MH, de Castro M, et al. Surrogate endpoints for prostate cancer-specific mortality after radiotherapy and androgen suppression therapy in men with localised or locally advanced prostate cancer: an analysis of two randomised trials. Lancet Oncol 2012;13:189-95.
33. Aref I, Eapen L, Agboola O, et al. The relationship between biochemical failure and time to nadir in patients treated with external beam therapy for T1-T3 prostate carcinoma. Radiother Oncol 1998;48:203-7.
34. Jackson WC, Johnson SB, Foster B, et al. Combining prostate-specific antigen nadir and time to nadir allows for early identification of patients at highest risk for development of metastasis and death following salvage radiation therapy. Pract Radiat Oncol 2014;4:99-107.
35. Ray ME, Levy LB, Horwitz EM, et al. Nadir prostate-specific antigen within 12 months after radiotherapy predicts biochemical and distant failure. Urology 2006;68:1257-62.
36. Johnson SB, Jackson WC, Murgic J, et al. Time to Nadir PSA: Of Popes and PSA--The Immortality Bias. Am J Clin Oncol 2015;38:465-71.
37. Taylor JM, Park Y, Ankerst DP, et al. Real-time individual predictions of prostate cancer recurrence using joint models. Biometrics 2013;69:206-13.
38. Consensus statement: guidelines for PSA following radiation therapy. American Society for Therapeutic Radiology and Oncology Consensus Panel. Int J Radiat Oncol Biol Phys 1997;37:1035-41.
39. Proust-Lima C, Taylor JM, Williams SG, et al. Determinants of change in prostate-specific antigen over time and its association with recurrence after external beam radiation therapy for prostate cancer in five large cohorts. Int J Radiat Oncol Biol Phys 2008;72:782-91.
40. Taylor JM, Yu M, Sandler HM. Individualized predictions of disease progression following radiation therapy for prostate cancer. J Clin Oncol 2005;23:816-25.
41. Proust-Lima C, Taylor JM. Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach. Biostatistics, 2009;10:535-49.
42. Taylor JM. Individualized Predictions of Disease Progression Following Radiation Therapy for Prostate Cancer. J Clin Oncol 2005;23:816-25.
43. Valicenti RK, DeSilvio M, Hanks GE, et al. Posttreatment prostatic-specific antigen doubling time as a surrogate endpoint for prostate cancer-specific survival: an analysis of Radiation Therapy Oncology Group Protocol 92-02. Int J Radiat Oncol Biol Phys 2006;66:1064-71.
44. Denham JW, Steigler A, Wilcox C, et al. Time to biochemical failure and prostate-specific antigen doubling time as surrogates for prostate cancer-specific mortality: evidence from the TROG 96.01 randomised controlled trial. Lancet Oncol 2008;9:1058-68.
45. Buyyounouski MK, Hanlon AL, Horwitz EM, et al. Interval to biochemical failure highly prognostic for distant metastasis and prostate cancer-specific mortality after radiotherapy. Int J Radiat Oncol Biol Phys 2008;70:59-66.
46. Buyyounouski MK, Pickles T, Kestin LL, et al. Validating the interval to biochemical failure for the identification of potentially lethal prostate cancer. J Clin Oncol 2012;30:1857-63.
47. Farr W. Vital statistics: memorial volume of selections from the reports and writings. 1885. Bull World Health Organ 2000;78:88-95.
48. Cavanaugh SX, Kupelian PA, Fuller CD, et al. Early prostate-specific antigen (PSA) kinetics following prostate carcinoma radiotherapy: prognostic value of a time-and-PSA threshold model. Cancer 2004;101:96-105.
49. Critz FA. Time to achieve a prostate specific antigen nadir of 0.2 ng./ml. after simultaneous irradiation for prostate cancer. J Urol 2002;168:2434-8.
50. Ray ME, Thames HD, Levy LB, et al. PSA nadir predicts

biochemical and distant failures after external beam radiotherapy for prostate cancer: a multi-institutional analysis. Int J Radiat Oncol Biol Phys 2006;64:1140-50.
51. Souhami L, Bae K, Pilepich M, et al. Impact of the duration of adjuvant hormonal therapy in patients with locally advanced prostate cancer treated with radiotherapy: a secondary analysis of RTOG 85-31. J Clin Oncol 2009;27:2137-43.
52. Williams SG. Characterization of the behavior of three definitions of prostate-specific antigen-based biochemical failure in relation to detection and follow-up biases: comparison with the American Society for Therapeutic Radiology and Oncology consensus definition. Int J Radiat Oncol Biol Phys 2006;64:849-55.
53. Williams SG. Ambiguities within the ASTRO consensus definition of biochemical failure: never assume all is equal. Int J Radiat Oncol Biol Phys 2004;58:1083-92.
54. Horwitz EM, Vicini FA, Ziaja EL, et al. The correlation between the ASTRO Consensus Panel definition of biochemical failure and clinical outcome for patients with prostate cancer treated with external beam irradiation. Int J Radiat Oncol Biol Phys 1998;41:267-72.
55. Vicini FA, Kestin LL, Martinez AA. The importance of adequate follow-up in defining treatment success after external beam irradiation for prostate cancer. Int J Radiat Oncol Biol Phys 1999;45:553-61.
56. Connell PP, Ignacio L, McBride RB, et al. Caution in interpreting biochemical control rates after treatment of prostate cancer: length of follow-up influences results. Urology 1999;54:875-9.
57. Pickles T, Kim-Sing C, Morris WJ, et al. Evaluation of the Houston biochemical relapse definition in men treated with prolonged neoadjuvant and adjuvant androgen ablation and assessment of follow-up lead-time bias. Int J Radiat Oncol Biol Phys 2003;57:11-8.
58. Pickles T; British Columbia Cancer Agency Prostate Cohort Outcomes Initiative. Prostate-specific antigen (PSA) bounce and other fluctuations: which biochemical relapse definition is least prone to PSA false calls? An analysis of 2030 men treated for prostate cancer with external beam or brachytherapy with or without adjuvant androgen deprivation therapy. Int J Radiat Oncol Biol Phys 2006;64:1355-9.
59. Critz FA, Williams WH, Benton JB, et al. Prostate specific antigen bounce after radioactive seed implantation followed by external beam radiation for prostate cancer. J Urol 2000;163:1085-9.
60. Stock RG, Stone NN, Cesaretti JA. Prostate-specific antigen bounce after prostate seed implantation for localized prostate cancer: descriptions and implications. Int J Radiat Oncol Biol Phys 2003;56:448-53.
61. Rosser CJ, Kuban DA, Levy LB, et al. Prostate specific antigen bounce phenomenon after external beam radiation for clinically localized prostate cancer. J Urol 2002;168:2001-5.
62. Crook J, Gillan C, Yeung I, et al. PSA kinetics and PSA bounce following permanent seed prostate brachytherapy. Int J Radiat Oncol Biol Phys 2007;69:426-33.
63. Williams SG, Pickles T, Kestin L, et al. A multicenter study demonstrating discordant results from electronic prostate-specific antigen biochemical failure calculation systems. Int J Radiat Oncol Biol Phys 2006;65:1494-500.
64. Sweeney CJ, Chen YH, Carducci M, et al. Chemohormonal Therapy in Metastatic Hormone-Sensitive Prostate Cancer. N Engl J Med 2015;373:737-46.
65. Fizazi K, Tran N, Fein L, et al. Abiraterone plus Prednisone in Metastatic, Castration-Sensitive Prostate Cancer. N Engl J Med 2017;377:352-60.
66. Beer TM, Armstrong AJ, Rathkopf DE, et al. Enzalutamide in metastatic prostate cancer before chemotherapy. N Engl J Med 2014;371:424-33.
67. Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. N Engl J Med 2012;367:1187-97.
68. Institute of Medicine. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Washington, DC: The National Academies Press, 2010.
69. Zhao F. Surrogate End Points and Their Validation in Oncology Clinical Trials. J Clin Oncol 2016;34:1436-7.
70. International conference on harmonisation; guidance on statistical principles for clinical trials; availability--FDA. Notice. Fed Regist 1998;63:49583-98.

Cite this article as: Williams S. Surrogate endpoints in early prostate cancer research. Transl Androl Urol 2018;7(3):472-482. doi: 10.21037/tau.2018.05.10