US National Cancer Institute investigates PSA coding errors

PSA biomarker data have not been included in this year’s release of data from the US National Cancer Institute (NCI)’s large Surveillance, Epidemiology, and End Results (SEER) database pending investigation into coding errors.

“In a study developed as part of our routine quality assurance and quality assessment activities in the SEER programme, we noted that a substantial number of registrar-reported PSA values were incorrect,” says Lynne Penberthy, associate director for the Surveillance Research Program (Division of Cancer Control and Population Sciences, NCI [Bethesda, MD, USA]).

The initial study involved re-abstraction of de-identified medical records from registrars in the USA. Subsequent preliminary assessments of PSA values in two regional SEER cancer registries (in Georgia and California), found an error rate that ranged from 5% to 15%. This value was lower than that reported in the initial study (about 18%)—probably because of extensive, ongoing quality-assurance activities at SEER registries, Penberthy says.

“We felt that it was imperative to remove PSA from our dataset until we could more comprehensively evaluate the PSA data in all our SEER registries and correct the errors”, Penberthy told The Lancet Oncology.

“The impact is immediate”, says Scott Williams (Peter MacCallum Cancer Centre [Melbourne, VA, Australia]). He stated, “I have several large scale analyses in process which have occupied many hundreds of hours of investigators’ time. These are all compromised now and I am reconsidering their viability.”

On April 15, 2015, SEER announced the discovery of the errors and the decision to remove PSA data from the present file. After that announcement, some journals are now rejecting research papers that use PSA values from SEER, reported Ronald C Chen (University of North Carolina [Chapel Hill, NC, USA]).

Chen postulates that, “many SEER papers on prostate cancer are perhaps on hold until the PSA issue is resolved”.

These errors stem from confusion about how to properly encode PSA values among cancer registrars at hospitals, cancer centres, and state cancer registries. “PSA is coded in a three-digit field with an implied decimal between the second and third digits”, explains Kevin Ward, principal investigator for the Georgia SEER registry (Atlanta, GA, USA). “For example, a PSA of 4·0 ng/mL should be coded as 040. Some registrars were confused with proper use of the implied decimal, resulting in coding a PSA of 4·0 ng/mL, for example, incorrectly as 004.”

SEER and contract firm WESTAT (Rockville, MD, USA) are developing a protocol to assess error rates throughout all SEER registries and to correct identified erroneous values of PSA “for recent years”, Penberthy says. “Corrected data will be reposted and made available to researchers.”

How much time that process will require is not yet clear. Penberthy explains that some registrars will need to, “follow back to the hospitals and other reporting sources”. She reassured The Lancet Oncology that “we [NCI] have identified this as a top priority in the SEER programme”, but added that corrections to PSA will take additional time.

Meanwhile, Penberthy commented that of “imperative” importance is the removal of the implied decimal field from PSA data-entry methods. Similarly, several other variables—including other biomarker values and tumour size—might require the registrar to adjust the decimal point. However, excluding tumour size, few of these are as integral to prostate cancer staging as PSA is, according to Penberthy. “We are looking at all the variables that might be impacted by the implied decimal [problem]. However, because of the high error rate in PSA this is our initial focus.”

Previous studies reported tumour size data in SEER have 95–99% accuracy, depending on cancer site, Penberthy notes.

PSA errors seem to have “minimal” statistical impact (3–4%) on cancer stage categorisations, suggest preliminary analyses in the original study, Penberthy commented.

However, subset analyses that use PSA values might be affected, cautions Williams. “If we use PSA as a stratification or matching variable in analyses, errors may impact strongly [on subset analyses]. Similarly, isolation of subsets with PSA would be problematic—such as doing analysis on ‘low-risk’ men. These men are defined, in part, as those with PSA values below 10 ng/mL. If miscoded, they would appear to be 1·4 instead of 14, so you could have quite erroneous data and conclusions from dealing with such subsets.”

“Until we have further information, it will be difficult to say how this might impact analyses using the PSA variable”, Penberthy says.

The PSA value errors should “not at all” affect the recommendations of ASCO’s Provisional Clinical Opinion regarding PSA screening decisions, says Chen.

SEER’s transparency about the errors is reassuring, agree Williams and Chen. “The identification of the problem is a testament to the rigour of validation processes within their massive database”, Williams says.

The North American Association of Central Cancer Registries (NAACCR) is working with NCI-SEER to address the PSA errors. “We will be developing new training materials and adapting our data standards as necessary”, says Betsy Kohler (NAACCR, Springfield, IL, USA). “We hope these issues will get resolved quickly and that the data will be available for use once again.”

Bryant Furlow

Published online May 1, 2015 http://dx.doi.org/10.1016/S1470-2045(15)70196-8
For more about the SEER programme and its national network of US regional cancer registries see http://seer.cancer.gov/registries/list.html
For the US National Cancer Institute SEER programme’s online announcement about errors in data for PSA see http://seer.cancer.gov/data/psa-values.html