The analysis of clinical cohort studies is a very important and valuable method performed to confirm and validate the results of cancer research, especially in those from basic biological studies. However, cohort studies require considerable effort, cost, and time for one researcher. To address such a situation, The Cancer Genome Atlas (TCGA) project was launched in 2005 and has published huge amounts of datasets in various cancer tissues including comprehensive information on mutations, gene expression, copy number variation, and DNA methylation. Since these data are linked to detailed clinical information of each case including the tissue type, clinical stage, therapeutic regimen, and survival duration, we can analyze the relationship between genomic and clinical information in various cancers. TCGA project was summarized and completed as the Pan-Cancer Atlas.1–3 Although these data are publicly available and we can obtain valuable information by performing statistical analyses of these data, not all researchers are good at treating and analyzing such data adequately. Recently, several websites have provided tools that are relatively easy to use to analyze TCGA data, and many researchers have taken advantage of such sites for their studies. Additionally, the figures in a paper are sometimes exactly the same as those generated by these sites. However, we often encounter the situations in which TCGA data that have been analyzed in papers we read or reviewed cannot be reproduced, even if TCGA datasets are used, especially in survival analyses. Therefore, we attempted to confirm the data source used for TCGA survival analyses.

First, we compared the overall survival data between 11,315 cases from the GDC data portal (https://portal.gdc.cancer.gov/) in which all TCGA data have been deposited and 11,160 cases from the paper of Pan-Cancer Atlas4 which is recommended as a resource of survival data for analysis. We found differences in days to death (death days) in 46 cases (increased except for one case) and in days to the last follow-up (follow days) in 56 cases (all increased) (Supplementary Table S1). Furthermore, the living status was changed in 24 cases. In 14 cases, the status was changed from dead to alive (Supplementary Table S2). If these differences in survival data are not adequately treated and processed, they may have a significant effect on survival analyses, resulting in misinterpretation.

We selected three widely used websites that provide tools that can be used to analyze survival data from TCGA, the Human Protein Atlas (https://www.proteinatlas.org/), KM plotter (https://kmplot.com/analysis/), and UCSC Xena (http://xena.ucsc.edu/). The provisional TCGA dataset in cBioPortal (https://www.cbioportal.org/) was also selected as a survey, although cBioPortal can adequately handle survival data from Pan-Cancer Atlas. We investigated 12 cases in whom the living status was changed from dead to alive with the information of days (Table 1). In most cases, day numbers were incorrectly imported. For example, the deaths of a total of six cases with urinary tract (BLCA), breast (BRCA), cervix (CESC), head and neck (HNSC), rectum (READ), and stomach (STAD) cancers were accompanied with follow days, not death days, in all three sites. In particular, the status change of one BRCA case (with 26 days) and one STAD case (with 21 days) may have significant effects on statistical analyses of survival because of the short-term deaths. In fact, the difference in survival data caused a significant change in the p-value (Supplementary figure). GEPIA (http://gepia.cancer-pku.cn/) also provides tools for survival analyses and many researchers use it. However, we were not able to verify the source for analysis because only graph figures, not text survival data from each case are included in the output of analysis. GEPIA provides not only overall survival (OS) but also disease-free
survival (DFS) data. Strangely, the case numbers of OS and DFS in GEPIA are exactly the same; nevertheless, TCGA provides DFS data on some cases. Thus, the analysis is "black box".

It goes without saying that the survival analyses using TCGA datasets have great power for cancer research. However, it demonstrates its true value as far as the data is treated and analyzed adequately, as a matter of course. When the results of survival analyses generated by web tools are evaluated, even though the analysis is based on TCGA datasets, the underlying source of survival data should be confirmed (rather than easily accepting the results as is), although it is advisable that each researcher performs the survival analysis themselves with raw survival data provided by TCGA.

**Disclosure of potential conflicts of interest**

No potential conflicts of interest were disclosed.

**References**

1. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V, et al. 2018. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 173(2):291–304 e296
2. Sanchez-Vega F, Mina M, Armenia J, Chhatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, et al. Oncogenic signaling pathways in The Cancer Genome Atlas. Cell. 2018;173(2):321–37 e10.
3. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. Pathogenic germ-line variants in 10,389 adult cancers. Cell. 2018;173:355–70 e14.
4. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatch AJ, Benz CC, Levine DA, Lee AV, et al. An integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173:400–16 e11.