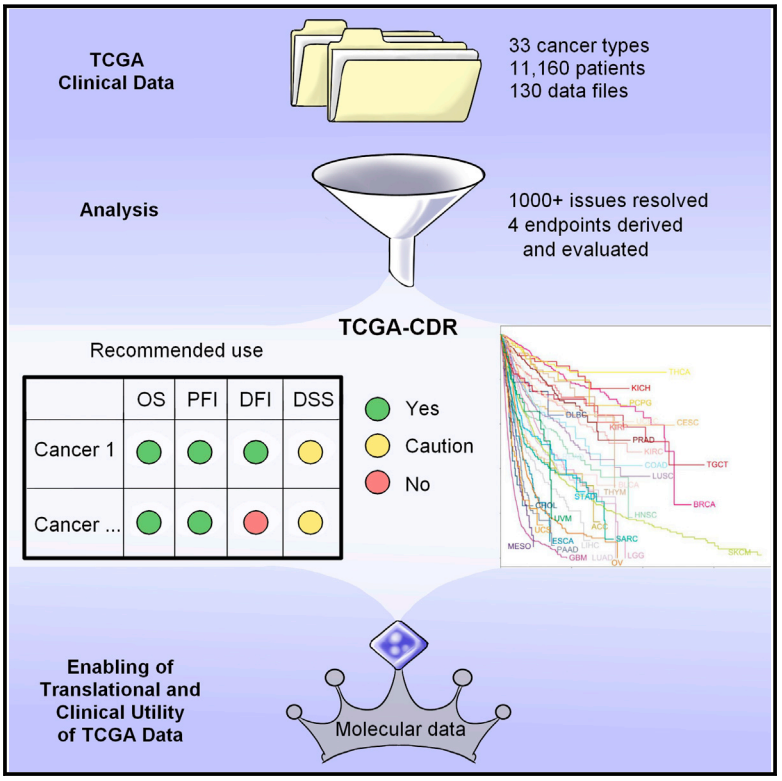
Article

Cell

An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics

Graphical Abstract



Highlights

- Generation of TCGA Clinical Data Resource for 11,160 patients over 33 cancer types
- Analysis of clinical outcome endpoints with usage recommendations for each cancer
- Demonstration of data validity and utility for large-scale translational research

Authors

Jianfang Liu, Tara Lichtenberg, Katherine A. Hoadley, ..., Vesteinn Thorsson, The Cancer Genome Atlas Research Network, Hai Hu

Correspondence

h.hu@wriwindber.org

In Brief

Analysis of clinicopathologic annotations for over 11,000 cancer patients in the TCGA program leads to the generation of TCGA Clinical Data Resource, which provides recommendations of clinical outcome endpoint usage for 33 cancer types.

Liu et al., 2018, Cell *173*, 400–416
April 5, 2018 © 2018 Elsevier Inc.
https://doi.org/10.1016/j.cell.2018.02.052

CellPress

Cell

Article

An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics

Jianfang Liu,[1] Tara Lichtenberg,[2] Katherine A. Hoadley,[3] Laila M. Poisson,[4] Alexander J. Lazar,[5] Andrew D. Cherniack,[6] Albert J. Kovatich,[7] Christopher C. Benz,[8] Douglas A. Levine,[9] Adrian V. Lee,[10] Larsson Omberg,[11] Denise M. Wolf,[12] Craig D. Shriver,[13] Vesteinn Thorsson,[14] The Cancer Genome Atlas Research Network, and Hai Hu[1,15,*]

[1]Chan Soon-Shiong Institute of Molecular Medicine at Windber, Windber, PA 15963, USA
[2]Nationwide Children's Hospital, Columbus, OH 43205, USA
[3]Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[4]Henry Ford Cancer Institute and Hermelin Brain Tumor Center, Henry Ford Health System, Detroit, MI 48202, USA
[5]Departments of Pathology, Genomic Medicine, and Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[6]The Eli and Edythe L. Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[7]Clinical Breast Care Project, Murtha Cancer Center, Uniformed Services University/Walter Reed National Military Medical Center, Bethesda, MD 20889, USA
[8]Buck Institute for Research on Aging, Novato, CA 94945, USA
[9]Division of Gynecologic Oncology, Department of OB/GYN, NYU Langone Medical Center, New York, NY 10016, USA
[10]Department of Pharmacology and Chemical Biology and Human Genetics, University of Pittsburgh, Women's Cancer Research Center, UPMC Hillman Cancer Center and Magee-Womens Research Institute, Pittsburgh, PA 15213, USA
[11]Sage Bionetworks, Seattle, WA 98109, USA
[12]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA 94143, USA
[13]Murtha Cancer Center, Uniformed Services University/Walter Reed National Military Medical Center, Bethesda, MD 20889, USA
[14]Institute for Systems Biology, Seattle, WA 98109, USA
[15]Lead Contact
*Correspondence: h.hu@wriwindber.org
https://doi.org/10.1016/j.cell.2018.02.052

SUMMARY

For a decade, The Cancer Genome Atlas (TCGA) program collected clinicopathologic annotation data along with multi-platform molecular profiles of more than 11,000 human tumors across 33 different cancer types. TCGA clinical data contain key features representing the democratized nature of the data collection process. To ensure proper use of this large clinical dataset associated with genomic features, we developed a standardized dataset named the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR), which includes four major clinical outcome endpoints. In addition to detailing major challenges and statistical limitations encountered during the effort of integrating the acquired clinical data, we present a summary that includes endpoint usage recommendations for each cancer type. These TCGA-CDR findings appear to be consistent with cancer genomics studies independent of the TCGA effort and provide opportunities for investigating cancer biology using clinical correlates at an unprecedented scale.

INTRODUCTION

The purpose of The Cancer Genome Atlas (TCGA) project was to establish a coordinated team science effort to comprehensively characterize the molecular events in primary cancers and to provide these data to the public for use by researchers around the world. TCGA started in 2006 with a 3-year pilot project focusing on glioblastoma multiforme (GBM), lung squamous cell carcinoma (LUSC), and ovarian serious cystadenocarcinoma (OV), followed by the execution of the full project from 2009 to 2015. By the end of this 10-year project, TCGA network investigators had characterized the molecular landscape of tumors from 11,160 patients across 33 cancer types and defined their many molecular subtypes. The quantity and quality of TCGA molecular data have been lauded by a large number of scientists, and these data have resulted in studies that have significantly advanced our understanding of cancer biology, as documented in dozens of highly cited published TCGA marker and companion papers, including those for GBM, OV, and breast, lung, prostate, bladder, and other individual cancers (Cancer Genome Atlas Network, 2012, 2015; The Cancer Genome Atlas Research Network, 2008, 2011, 2012, 2014, 2015; Cancer Genome Atlas Research Network et al., 2017). TCGA data also make possible studies that compare and contrast multiple cancer types with the goal of identifying common themes that transcend the tissue of origin and may inform precision oncology (Hoadley et al., 2014). In addition, numerous independent investigators have used TCGA as a resource to support their own studies and to help interpret molecular testing of individual patients in a clinical setting (Huo et al., 2017; Verhaak et al., 2010). However, obtaining comprehensive clinical annotation was neither a primary program objective nor a practical possibility, given the worldwide scope and severe time constraints for sample accrual goals

 © 2018 Elsevier Inc.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

determined at the time of TCGA program initiation and funding. The incomplete annotation of patient outcome and treatment data associated with each TCGA-acquired sample, with its relatively short-term clinical follow-up interval, has been noted by the research community (Hoadley et al., 2014; Huo et al., 2017). The limitations of the existing clinical dataset, associated with an otherwise rich body of genomic and molecular analyses available across all TCGA tumor types, compels thorough and systematic curation and evaluation of those clinical endpoints and other clinical features associated with each TCGA tumor so that the scientific community can optimize the translational relevance of the tumor-specific genomic and pathway conclusions drawn from the TCGA program and its pan-cancer analyses. It is also important to demonstrate that the conclusions drawn from this newly curated TCGA pan-cancer clinical data resource have translational validity with respect to both patient prognosis and outcome parameters.

In clinical studies, 5-year or 10-year benchmark survival rates are often calculated to convey prognostic information or to compare treatment effects. These survival rates may be based on progression or mortality events with or without disease specificity. For each endpoint, it is very important to have a sufficiently long follow-up time to capture the events of interest, and the minimum follow-up time needed depends on both the aggressiveness of the disease and the type of endpoint (Tai et al., 2005).

Overall survival (OS) is an important endpoint, with the advantage that there is minimal ambiguity in defining an OS event (Hudis et al., 2007; Punt et al., 2007); the patient is either alive or dead. However, using OS as an endpoint may weaken a clinical study as deaths because of non-cancer causes do not necessarily reflect tumor biology, aggressiveness, or responsiveness to therapy. Using OS or disease-specific survival (DSS) demands longer follow-up times; thus, in many clinical trials, disease-free interval (DFI) or progression-free interval (PFI) are used (Hudis et al., 2007; Punt et al., 2007; https://wiki.nci.nih.gov/plugins/servlet/mobile#content/view/24279961). The minimum follow-up time for these endpoints is shorter because patients generally develop disease recurrence or progression before dying of their disease. Selection of a specific survival endpoint also depends on the study goal. For example, a clinical trial testing the effect of a drug's ability to delay or prevent cancer progression would use PFI as the most appropriate endpoint. With specific regard to the analysis of available TCGA clinical data, it is important to realize that short-term clinical follow-up intervals favor outcome analyses in more aggressive cancer types, which are likely to observe events within a couple of years. Studies with less aggressive cancer types, in which patients relapse only after many years or even decades, may not observe enough events during their follow-up intervals to support reliable outcome determinations. The intent of this analysis is to examine the relative strengths and weaknesses of the TCGA pan-cancer clinical outcome measures to guide future analyses and avoid pitfalls such as insufficient follow-up intervals.

To our knowledge, there has been no systematic attempt to analyze the TCGA clinical data and derive acceptable outcome endpoints across all 33 TCGA cancer types involving 11,160 patients or to assess the adequacy of the clinical follow-up interval for each survival endpoint test. Here we present curated and filtered clinical and survival outcome data as a newly integrated resource for the entire scientific community, describe how problems encountered while analyzing these data were resolved, and what pitfalls researchers should be aware of when using these data for future correlative and survival studies. Based on our comprehensive clinical review, we also provide scoring recommendations for appropriate future use and tumor-specific endpoint selection. The resulting compendium of curated data is now presented as the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) for public access and future translational cancer research.

RESULTS

The TCGA clinical data were downloaded from the data portal of the Genomic Data Commons (GDC), where all TCGA molecular data are also available (https://gdc-portal.nci.nih.gov/legacy-archive/). The same TCGA barcode structure is used for both clinical data and molecular data, enabling integrated analysis of patient-based clinical data and sample-based molecular data.

Cohort Characteristics

Figure 1A shows a flowchart of the methods for clinical data integration and analysis as well as derivation and evaluation of 4 major clinical outcome endpoints. We processed 33 initial enrollment data files and 97 follow-up data files for 11,160 patients across 33 cancer types. Table 1 shows the basic characteristics of each TCGA cohort. Primary tumor samples, not metastatic, were typically selected in each cohort for molecular characterization, with the exception of the skin cutaneous melanoma (SKCM) study, which allowed both. A very limited number of metastatic tumors with matching primary tumors was also studied for other cancer types. Individual patients' detailed data are provided in Table S1, tab TCGA-CDR, and problems we identified when processing this dataset and the solutions we developed are described in the STAR Methods.

Clinical Outcome Endpoints of OS, PFI, DFI, and DSS

There are many definitions of clinical outcomes used in oncology research. After analyzing all TCGA clinical data used for this study, we concluded that OS and PFI could be derived relatively accurately using the available data. We also derived DFI reasonably accurately, although, for most cases, DSS could only be estimated. Figure 1B shows the OS Kaplan-Meier (K-M) plots for all cases of the 33 different cancer types. Although TCGA did not set survival analyses as a primary program objective, the resulting survival plots for most cancer types are similar to prior independent studies prospectively designed to evaluate these same survival endpoints. This is perhaps best exemplified by the TCGA outcomes for GBM, OV (Cancer Genome Atlas Research Network, 2008, 2011), and lower-grade glioma (LGG) (Cancer Genome Atlas Research Network et al., 2015). K-M plots for PFI, DFI, and DSS are shown in Figures 1C–1E (see also Figure S1).

We calculated median follow-up times as well as median times to event or censorship based on the observed times for these four endpoints for each cancer type (Table 2; Table S1, tab

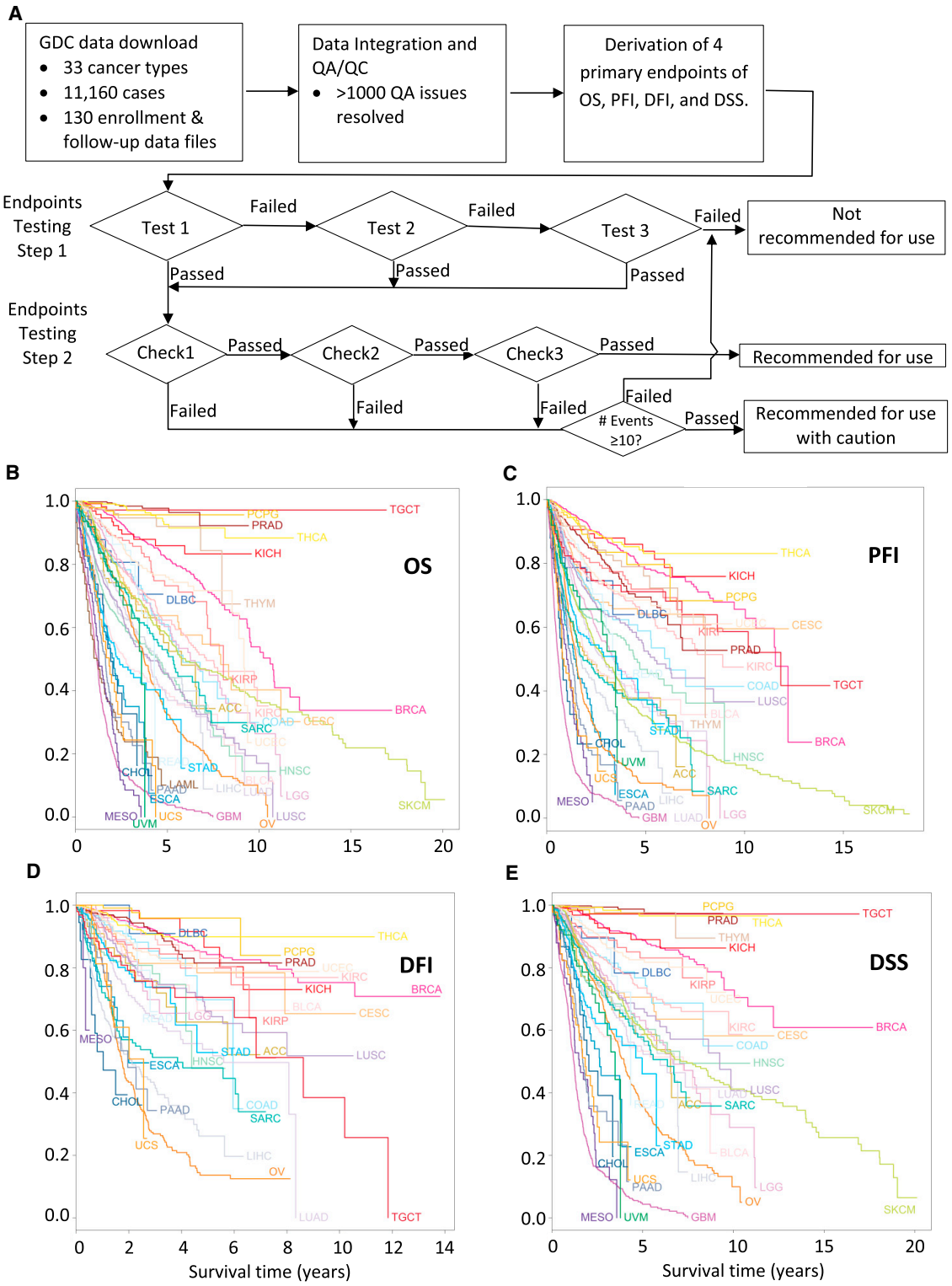


Figure 1. Clinical Data Analysis

(A–E) Flowchart (A) and K-M plots of the Pan-Cancer types for OS (B), PFI (C), DFI (D), and DSS (E) respectively. The tail of each K-M curve is truncated at the point when fewer than 10 patients remain at risk. See Figure S1 for plots of the 4 endpoints within each of the 33 tumor types and Tables S1, S2, S3 and S4 for more detailed information regarding endpoint derivation.

Table 1. TCGA Pan-Cancer Cohort Characteristics

Cancer Type	No. of Cases	Age[a] (Mean ± SD)	Gender M/F	Race White/Black/Other/NA	Stage[b] 0/I/II/III/IV/NA	Grade[c] 1/2/3/4/NA
ACC	92	47.2 ± 16.3	32/60	78/1/2/11	0/9/44/19/18/2	0/0/0/0/92
BLCA	412	68.1 ± 10.6	304/108	327/23/44/18	0/2/131/141/136/2	21/0/388/0/3
BRCA	1097	58.4 ± 13.2	12/1085	757/183/62/95	0/183/621/249/20/24	0/0/0/0/1097
CESC	307	48.3 ± 13.8	0/307	211/30/30/36	0/163/70/46/21/7	18/136/120/1/32
CHOL	45	63.6 ± 12.2	20/25	38/3/3/1	0/20/11/4/10/0	1/22/20/2/0
COAD	459	66.9 ± 13.1	243/216	214/59/12/174	0/76/178/129/65/11	0/0/0/0/459
DLBC	48	56.3 ± 13.9	22/26	29/1/18/0	0/8/17/5/12/6	0/0/0/0/48
ESCA	185	62.5 ± 11.9	158/27	114/5/46/20	0/18/79/56/9/23	19/77/49/0/40
GBM	596	57.8 ± 14.4	366/230	507/51/13/25	0/0/0/0/0/596	0/0/0/596/0[d]
HNSC	528	60.9 ± 11.9	386/142	452/48/13/15	0/27/74/82/270/75	63/311/125/7/22
KICH	113	51.2 ± 13.9	62/51	95/12/4/2	0/54/33/19/7/0	0/0/0/0/113
KIRC	537	60.6 ± 12.2	346/191	466/56/8/7	0/269/57/125/83/3	14/230/207/78/8
KIRP	291	61.5 ± 12.1	214/77	207/61/8/15	0/173/21/52/15/30	0/0/0/0/291
LAML	200	55.0 ± 16.1	109/91	181/15/2/2	0/0/0/0/0/200	0/0/0/0/200
LGG	515	42.9 ± 13.4	285/230	475/21/9/10	0/0/0/0/0/515	0/249/265/0/1
LIHC	377	59.5 ± 13.5	255/122	187/17/163/10	0/175/87/86/5/24	55/180/124/13/5
LUAD	522	65.3 ± 10.0	242/280	393/53/9/67	0/279/124/85/26/8	0/0/0/0/522
LUSC	504	67.3 ± 8.6	373/131	351/31/9/113	0/245/163/85/7/4	0/0/0/0/504
MESO	87	63.0 ± 9.8	71/16	85/1/1/0	0/10/16/45/16/0	0/0/0/0/87
OV	587	59.7 ± 11.5	0/587	498/34/24/31	0/17/30/446/89/5	6/69/495/1/16[e]
PAAD	185	64.9 ± 11.1	102/83	162/7/11/5	0/21/152/4/5/3	32/97/51/2/3
PCPG	179	47.3 ± 15.1	78/101	148/20/7/4	0/0/0/0/0/179	0/0/0/0/179
PRAD	500	61.0 ± 6.8	500/0	147/7/2/344	0/0/0/0/0/500	0/0/0/0/500
READ	170	64.5 ± 11.9	92/78	82/6/1/81	0/33/51/52/25/9	0/0/0/0/170
SARC	261	60.9 ± 14.7	119/142	228/18/6/9	0/0/0/0/0/261	0/0/0/0/261
SKCM[f]	470	58.2 ± 15.7	290/180	447/1/12/10	7/77/140/171/23/52	0/0/0/0/470
STAD	443	65.7 ± 10.8	285/158	278/13/90/62	0/59/130/183/44/27	12/159/263/0/9
TGCT	134	32.0 ± 9.3	134/0	119/6/4/5	0/101/12/14/0/7	0/0/0/0/134
THCA	507	47.3 ± 15.8	136/371	334/27/53/93	0/285/52/113/55/2	0/0/0/0/507
THYM	124	58.2 ± 13.0	64/60	103/6/13/2	0/38/61/15/8/2	0/0/0/0/124
UCEC	548	63.9 ± 11.1	0/548	374/109/33/32	0/342/52/124/30/0	99/122/327/0/0
UCS	57	69.7 ± 9.3	0/57	44/9/3/1	0/22/5/20/10/0	0/0/0/0/57
UVM	80	61.6 ± 13.9	45/35	55/0/0/25	0/0/39/36/4/1	0/0/0/0/80

ACC, adrenocortical carcinoma; CHOL, cholangiocarcinoma; KIRC, kidney renal clear cell carcinoma; LIHC, liver hepatocellular carcinoma; READ, rectum adenocarcinoma; TGCT, testicular germ cell tumor; THCA, thyroid carcinoma; UCS, uterine carcinosarcoma; NA, not applicable.

[a]51 cases are missing age at diagnosis, and 46 patients were 90 years of age or older and were capped at 90 years because of Health Insurance Portability and Accountability Act (HIPAA) regulations.

[b]Included AJCC stage for most cancer types; clinical stages for CESC, DLBC, OV, UCEC, and UCS; and Masaoka stage for THYM. In the detailed data file shown in Table S1, tab TCGA-CDR, all originally reported stage types were retained.

[c]In BLCA, G1 was for "low grade" and G3 for "high grade" in this table; UCEC had 11 high grade, which was converted to G3 (the highest for this disease) in this table. All original values were retained in Table S1, tab TCGA-CDR.

[d]GBM is grade IV by definition. In the original TCGA dataset, the grades for GBM cases were not provided.

[e]We realized that OV should not have a grade IV disease but reported the data as in the original TCGA dataset.

[f]For SKCM, the majority of tumors were from bulky regional lymph node metastases or distant metastases, and the patients' initial diagnosis years of non-metastatic diseases, including stage 0 disease, were provided (*in situ*; see STAR Methods). No other cancer types had a stage 0 cancer diagnosis.

TCGA-CDR). The overall median follow-up time for all tumors was 22.1 months, but these times were very different across cancer types; GBM and acute myeloid leukemia (LAML) had the shortest (~12 months), whereas kidney chromophobe (KICH) had the longest (~48 months).

Recommended Use of Clinical Outcome Endpoints

Selection of the clinical outcome endpoints for a specific study depends on the goals of the study, number of events, cohort size, and quality of the outcome data. Methods are available to assess the quality of survival outcome data (Maller and Zhou,

Table 2. Median Follow-Up Times Overall and the Median Time to Event and to Censor for the Four Clinical Outcome Endpoints

Cancer Type	Median Follow-Up Time (Months)	OS Median Time (Months)		PFI Median Time (Months)		DFI Median Time (Months)		DSS Median Time (Months)	
	All	To Event	To Censor	To Event	To Censor	To Event	To Censor	To Event	To Censor
ACC	38.9	18.1	47.8	8.1	49.2	20.0	61.0	18.0	47.8
BLCA	17.6	13.5	21.0	9.7	17.8	14.8	19.1	13.6	19.4
BRCA	27.7	41.8	25.0	26.0	25.0	25.4	25.0	32.6	26.0
CESC	20.9	19.9	22.6	13.6	21.7	15.9	28.3	18.0	23.0
CHOL	21.6	18.0	30.1	7.1	22.4	7.1	25.3	18.3	28.3
COAD	22.0	13.3	24.0	12.0	22.5	16.0	29.3	11.1	24.0
DLBC	26.7	19.5	31.1	10.3	29.2	113.7	31.4	16.2	29.2
ESCA	13.1	11.5	13.2	8.8	12.6	7.4	13.2	13.5	12.9
GBM	12.0	12.6	8.5	6.1	5.9	31.5	26.3	12.7	8.4
HNSC	21.2	14.1	27.9	9.4	25.7	7.6	27.5	13.5	25.9
KICH	48.3	24.3	54.2	11.9	49.7	52.7	39.6	28.1	51.0
KIRC	39.0	26.9	47.8	13.5	43.0	29.6	45.4	23.2	46.4
KIRP	25.2	21.1	25.4	11.0	25.5	15.5	25.4	14.2	26.0
LAML	12.0	9.0	23.0	NA	NA	NA	NA	NA	NA
LGG	22.1	26.7	20.7	15.3	18.7	19.6	20.1	25.5	20.7
LIHC	19.7	13.7	21.3	9.0	15.6	9.0	17.6	19.8	19.7
LUAD	21.6	20.3	22.0	14.4	20.0	15.7	22.5	19.9	21.8
LUSC	21.9	18.1	24.9	14.0	21.1	18.0	26.9	18.8	22.5
MESO	16.9	15.0	38.4	10.3	19.4	15.5	9.8	14.7	24.9
OV	33.0	35.3	27.7	14.7	14.9	17.9	26.5	35.3	28.9
PAAD	15.3	12.9	17.0	11.2	13.8	14.8	15.7	13.8	15.9
PCPG	24.8	14.9	25.2	19.9	23.8	27.3	24.5	17.5	25.0
PRAD	30.5	36.2	30.5	18.4	28.2	24.9	30.4	43.7	30.5
READ	20.0	22.0	20.0	19.0	19.0	27.8	21.0	20.0	20.0
SARC	31.1	21.3	35.9	10.1	32.7	11.2	36.5	22.6	34.9
SKCM	35.9	35.3	36.9	23.5	22.7	21.8	23.8	36.6	34.3
STAD	14.0	11.3	17.2	9.5	13.8	10.8	18.6	12.4	16.1
TGCT	41.4	18.6	41.6	9.1	35.6	14.8	31.8	16.9	41.7
THCA	31.1	33.5	31.0	16.0	30.9	16.2	31.9	33.5	31.0
THYM	41.2	28.0	41.6	25.2	41.2	30.8	42.1	54.9	41.2
UCEC	29.9	23.3	31.2	16.8	29.7	17.1	30.7	21.9	30.7
UCS	20.1	17.1	27.2	9.0	26.9	16.6	27.2	14.7	26.9
UVM	25.8	19.9	27.0	12.5	25.0	12.2	26.2	19.9	27.0

1994; Shen, 2000). We applied these methods, and others we developed, as tests 1–3 and a supplemental check, to this dataset of individual diseases. We provided recommendations regarding how each outcome's endpoints should be used within each disease type, with concerns justified in comments (Table 3). Survival endpoints for each cancer type that passed at least one of the main tests as well as the supplemental check were considered acceptable for use. Overall, we recommend use of all four endpoints without reservation for 13 of the 33 cancer types: bladder urothelial carcinoma (BLCA), cervical squamous cell carcinoma (CESC), colon adenocarcinoma (COAD), esophageal carcinoma (ESCA), head and neck squamous cell carcinoma (HNSC), kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), LUSC, OV, pancreatic adenocarcinoma (PAAD), sarcoma (SARC), stomach adenocarcinoma (STAD), and uterine corpus endometrial carcinoma (UCEC). In contrast, none of the four outcome endpoints can be recommended for use in the TCGA pheochromocytoma and paraganglioma (PCPG) cases. For lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), LAML, and thymoma (THYM), only one endpoint can be recommended for use; for all other cancer types, either two or three endpoints can be recommended, some with particular reservations. Generally, the most reliable of all four endpoints, PFI, could be recommended for use without reservation in all but 4 of the 33 cancer types, the 4 exceptions being LAML (no data), DLBC and KICH (use with caution), and PCPG (not recommended). Thus, despite the impression held by many that TCGA follow-up times are too short for meaningful

Table 3. Assessment and Recommended Use of the Endpoints of OS, PFI, DFI, and DSS

		OS (Accurately Defined)			PFI (Accurately Defined)			DFI (Accurately Defined)			DSS (Approximately Defined)			
Type	N	Use	Event	Censored	Use	Event	Censored	Use	Event	Censored	Use	Event	Censored	Explanation/Caution
ACC	92	√	34	58	√	49	43	√*	14	39	√ app.	30	60	number of events is small
BLCA	412	√	181	231	√	177	235	√	32	157	√ app.	124	274	
BRCA	1097	√*	151	946	√	145	952	√	84	869	√ App.*	83	995	need a longer follow-up for OS and DSS
CESC	307	√	71	236	√	71	236	√	26	150	√ acc.	54	249	
CHOL	45	√	22	23	√	23	22	√*	10	18	√ app.*	18	24	sample size is too small for OS, DSS, DFI, and PFI
COAD	459	√	102	357	√	123	336	√	24	166	√ app.	64	379	
DLBC	48	X	9	39	√*	12	36	X	4	24	X	4	44	sample size and number of events are too small, need a longer follow-up
ESCA	185	√	77	108	√	87	98	√	23	66	√ app.	51	132	
GBM	596	√	491	105	√	506	90	X	2	1	√ app.	445	110	number of disease-free cases is small
HNSC	528	√	223	305	√	198	330	√	28	106	√ app.	130	372	
KICH	113	√*	13	100	√*	17	96	X	6	65	√ app.*	10	103	number of events is too small, need a longer follow-up
KIRC	537	√	177	360	√	162	375	√*	15	102	√ app.	110	415	number of events is small
KIRP	291	√	44	247	√	58	233	√	28	156	√ app.	28	259	
LAML	200	√	133	67	NA	NA	NA	NA	NA	NA	NA	NA	NA	only has OS data
LGG	515	√*	125	390	√	192	323	√	20	114	√ app.*	113	394	need a longer follow-up for OS and DSS
LIHC	377	√	132	245	√	185	192	√	147	176	√ app.*	80	288	need a longer follow-up DSS
LUAD	522	√	188	334	√	213	309	√	92	217	√ app.	116	370	
LUSC	504	√	219	285	√	149	355	√	63	241	√ app.	91	361	
MESO	87	√	74	13	√	61	26	X	7	8	√ app.	43	23	sample size for DFI is small
OV	587	√	349	236	√	414	173	√	196	90	√ app.	302	246	
PAAD	185	√	100	85	√	110	75	√	23	46	√ acc.	79	99	
PCPG	179	X	6	173	X	21	158	X	4	156	X	4	175	need a longer follow-up for OS, DSS, DFI, and PFI; number of events is small
PRAD	500	√*	10	490	√	93	407	√	30	310	X	5	493	need a longer follow-up for OS and DSS
READ	170	√*	26	144	√	39	131	X	7	41	√ app.*	15	149	need a longer follow-up for OS, DSS, and DFI; number of events for DFI is too small
SARC	261	√	99	162	√	139	122	√	67	86	√ app.	81	174	
SKCM	470	√	216	247	√	309	154	NA	NA	NA	√ app.	190	267	no information to derive DFI
STAD	443	√	172	271	√	143	300	√	46	213	√ app.	103	313	
TGCT	134	X	4	130	√	35	99	√	27	78	X	3	131	number of events is small for OS and DSS; need a longer follow-up

(*Continued on next page*)

Table 3. *Continued*

Type	N	OS (Accurately Defined)			PFI (Accurately Defined)			DFI (Accurately Defined)			DSS (Approximately Defined)			Explanation/Caution
		Use	Event	Censored	Use	Event	Censored	Use	Event	Censored	Use	Event	Censored	
THCA	507	√*	16	491	√	52	455	√	26	332	X	7	494	number of events is small for OS and DSS; need a longer follow-up
THYM	124	X	9	115	√	22	102	NA	NA	NA	X	4	120	number of events is too small for OS and DSS; need a longer follow-up; no information to derive DFI
UCEC	548	√	91	457	√	124	424	√	57	369	√ app.	60	486	
UCS	57	√	35	22	√	37	20	√*	10	17	√ app.	31	24	sample size is small
UVM	80	√	23	57	√	30	50	NA	NA	NA	√ acc.	21	59	no information to derive DFI

√, recommended for use (passed at least passed one of the 3 tests in step 1 and the supplemental checks in step 2 as described in the STAR Methods); X, not recommended for use; *, caution, see the explanation/caution column; app., approximate; acc., accurate.

endpoint analyses, in fact, they are sufficiently long for many endpoint determinations in the more aggressive tumor types and for determination of PFI in most tumor types, where disease progression events occur well before death events. The Cumulative event plots of OS, DSS, DFI, and PFI for each of 33 tumor types are provided in Figure S2.

Validation and Application Examples

In breast cancer studies, patients with estrogen receptor-negative (ER−) tumors have worse clinical survival outcomes compared with those with ER-positive (ER+) tumors. To evaluate the derived clinical endpoints, we compared the survival of patients with these two types of tumors using OS, PFI, DFI, and DSS, respectively (Figures 2A–2D; plots truncated at 10-year follow-up time, but analyses were conducted using the whole dataset following Huo et al., 2017). Univariate analyses showed that TCGA breast cancer patients with ER+ tumors had better survival than patients with ER− tumors when using PFI ($p = 0.005$) and DFI ($p = 0.001$) as clinical endpoints, but there was no sufficient evidence of a difference when using OS as the endpoint ($p = 0.097$). We also noticed that there was a significant difference in (approximated) DSS ($p = 0.009$), demonstrating the potential value of this estimated endpoint. As noted in Table 3, although we caution against using breast invasive carcinoma (BRCA) data to determine OS and DSS, the above findings validate our recommended use of PFI and DFI as suitable endpoints for specific types of breast cancer molecular studies.

We also examined the survival outcome endpoints of a more aggressive cancer type, GBM. The TCGA GBM median OS was 12.6 months, which falls between the previously reported 12.1 months with standard care and 14.6 months with standard care plus temozolomide (Stupp et al., 2005). The median PFI was 6.1 months, which falls between the reported 5.0 months with standard care and 6.9 months with standard care plus temozolomide (Stupp et al., 2005). Thus, the event time for OS and PFI derived from this TCGA dataset is consistent with the literature, an observation previously noted for OS in the initial GBM marker paper when only 185 cases were analyzed (Cancer Genome Atlas Research Network, 2008). This example again confirms the validity of OS and PFI as recommended clinical endpoints for correlation with GBM molecular studies.

We validate the curated TCGA-CDR data by using Cox proportional hazards regression models to determine the hazard ratio (HR) for patients with high-stage (III, IV) disease relative to patients with low-stage (I, II) disease for each of the four endpoints. Tumor-specific American Joint Committee on Cancer (AJCC) pathology stages (Amin et al., 2017) were employed following the then-current version used at each tissue source site (TSS). Because the definition of DFI was not consistent with that of other outcomes (i.e., cases with a follow-up of less than 90 days were excluded, so were *de novo* stage IV cases), we compared the logHRs of the stage-based measurements using the three other endpoints (OS, PFI, and DSS) for the 14 cancer types in which these outcome endpoints were recommended for use (Table 3). Subsequent statistical analyses using only diseases satisfying the Cox proportional hazards assumption (Grambsch and Therneau, 1994) were performed (Figures 2E–2G; Table S1, tab Figure 2EFG_AdditionalInfo). Our results showed that the high-stage HR was significantly larger than unity for most of the 14 cancer types and across the three recommended endpoints, with the exception of mesothelioma (MESO), PAAD, and uveal melanoma (UVM), which were not significantly different for high- versus low-stage disease for either OS, PFI, or DSS. K-M plots for these analyses are provided in Figure S3. Using paired Wilcoxon signed-rank test, the logHRs were significantly different when measured by PFI versus DSS ($p = 0.0008$) or PFI versus OS ($p = 0.039$), indicating evidence of a systematic difference in HR between the progression and survival endpoints. There was not significant evidence of a systematic difference between OS versus DSS ($p = 0.106$). Using Pearson correlation coefficients weighted by the inverse mean of two standard errors of the logHR values, we observed very strong positive associations between logHR estimates for all three outcomes: the correlation coefficient was 0.96 (95% confidence interval [CI]: 0.77–0.99) between PFI and OS, 0.95 (95% CI: 0.76–0.99) between PFI and DSS, and 0.90 (95% CI: 0.61–0.98) between OS and DSS. Notably, these

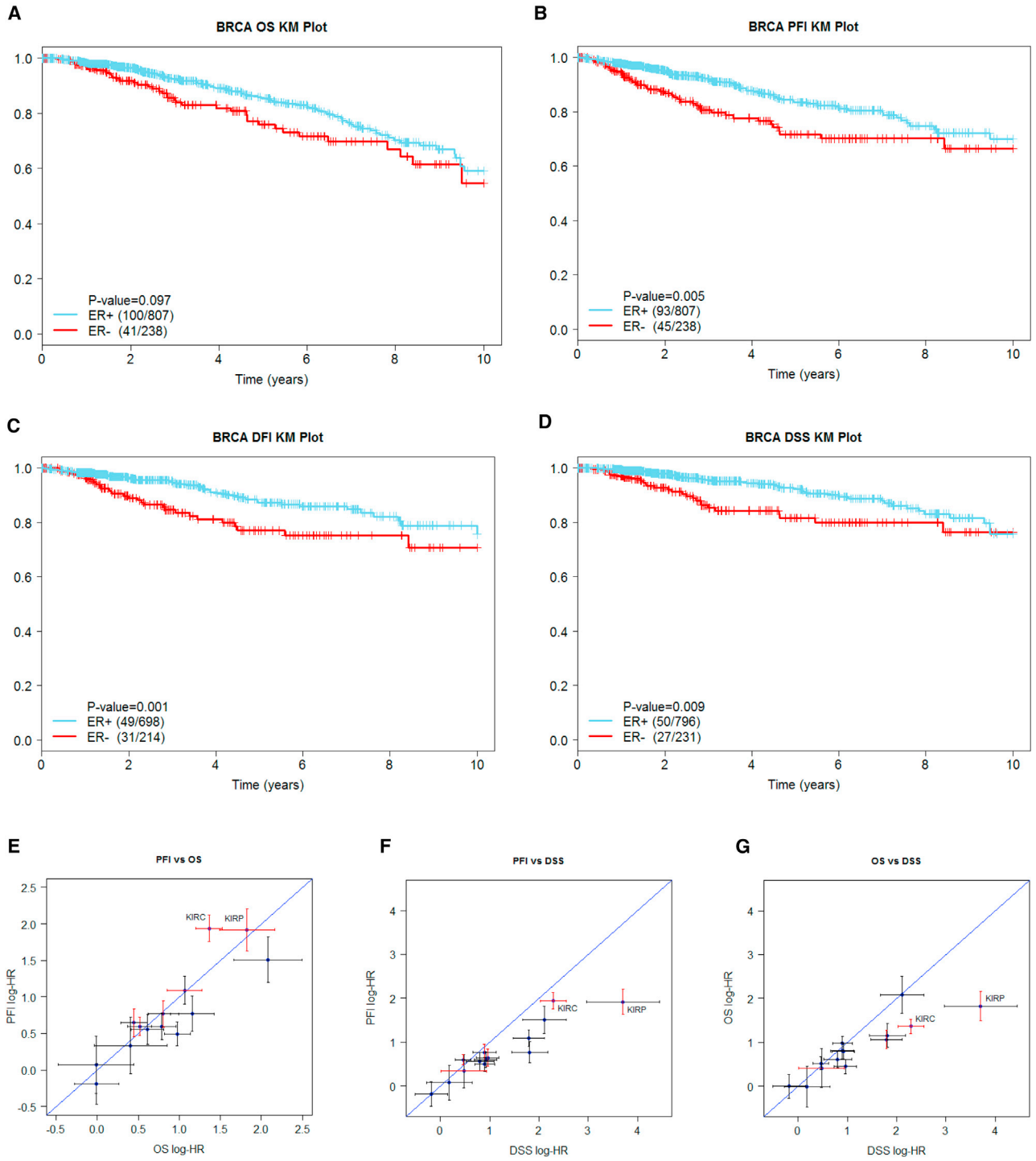


Figure 2. Validation and Application Examples

(A–D) Clinical survival outcomes of breast cancer patients with ER+ or ER− tumors using (A) OS, (B) PFI, (C) DFI, and (D) DSS as the endpoint, respectively. Plots were truncated at 10 years, but the analyses were conducted using all of the data.

(E–G) Pairwise plots comparing natural logHRs for event development measured by three different clinical endpoints (OS, DSS, and PFI), with natural logHRs calculated for high-stage (III, IV) disease relative to low-stage (I, II) disease: (E) PFI versus OS, (F) PFI versus DSS, and (G) OS versus DSS. Here 14 diseases for which these three endpoints are recommended for use are analyzed. The blue diagonal lines illustrate equal logHRs measured by each endpoint pair. Error bars represents the standard errors of logHR. Red error bars show models that did not satisfy the PHs assumption.

Table 4. NTE Development from Patients Who Were Never Disease-Free Compared with Those Who Were Once Disease-Free

		Disease-Free[a]		With Disease			
Type	No.	No.[c]	Event (n)	No.[c]	Event (n)	HR[b]	95% CI[b]
ACC	92	52	13	11	10	14.48	5.59–37.5*
BLCA	412	177	32	49	40	9.68	5.98–15.67*
BRCA	1097	890	80	31	4	1.81	0.66–4.95
CESC	307	172	26	16	7	4.7	2.02–10.94*
CHOL	45	23	7	*4*	2	2.31	0.48–11.19
COAD	459	187	22	24	15	10.13	5.19–19.74*
DLBC	48	27	4	*6*	3	21.99	2.25–215.2*
ESCA	185	78	17	13	10	3.84	1.75–8.41*
GBM	596	*3*	2	17	14	6.62	0.83–52.61**
HNSC	528	132	28	*9*	7	9.47	4.01–22.41*
KICH	113	68	6	*5*	2	15.61	2.17–112.59*
KIRC	537	112	15	*9*	4	*2.86*	*0.94–8.73*
KIRP	291	175	24	*7*	4	5.96	2.04–17.45*
LGG	515	124	20	292	136	3.12	1.95–5*
LIHC	377	277	131	16	11	1.8	0.97–3.34
LUAD	522	296	90	85	59	3.96	2.83–5.53*
LUSC	504	289	63	41	28	6.68	4.25–10.51*
MESO	87	14	7	*9*	7	1.64	0.57–4.71
OV	587	284	196	91	79	3.78	2.86–5*
PAAD	185	65	22	54	46	4.47	2.65–7.55*
PCPG	179	140	4	*9*	7	21.13	5.85–76.29*
PRAD	500	338	28	96	50	8.47	5.32–13.49*
READ	170	47	7	*5*	4	7.28	1.91–27.78*
SARC	261	146	61	69	40	1.92	1.28–2.86*
STAD	443	236	42	64	49	8.78	5.7–13.53*
TGCT	134	100	23	17	3	*0.5*	*0.15–1.7*
THCA	507	351	25	42	7	2.46	1.06–5.69*
UCEC	548	404	55	21	13	6.95	3.78–12.76*
UCS	57	25	10	11	11	20.91	5.58–78.29*
Pan-cancer	11,160	5,232	1,060	1,123	672	4.47	4.06–4.93*

*, FDR-adjusted q value < 0.05; **, model with a nominal p value from log rank test < 0.05. *Italic HR and 95% CI indicate* models we recommend for additional testing of PHs assumptions (c.f. Table S1, tab Table4_PHAssumptionTests).

[a]To overcome immortal time bias, we restricted the patient set to those surviving at least 3 months. This provides a proxy for the time required for a patient to complete treatment and be identified as disease-free.

[b]HR and 95% CI were calculated from Cox PHs regression models, using disease-free as reference.

[c]Italic and underscored, number of patients at risk < 10.

correlations support the potential clinical use of earlier-determined PFI as a proxy for later-determined OS and DSS endpoints (Shi and Sargent, 2009).

Apart from integration with the molecular data, there are many ways to use the pan-cancer data in this TCGA-CDR. As one application of the TCGA-CDR, we looked to see whether the likelihood of developing a new tumor event (NTE) differed between patients who were disease-free relative to those not disease-free following primary therapy. Twenty-nine cancer types in our TCGA-CDR had the information for addressing this question. For this analysis, we included patients who survived at least 3 months from diagnosis to approximate the time patients required to complete primary therapy and achieve a disease free state (i.e., to prevent immortal time bias in the disease-free group; Anderson et al., 1983; Giobbie-Hurder et al., 2013). Using LUSC as an example, there were 289 disease-free cases and 41 never disease-free cases with a 21.8% and 68.2% NTE rate, respectively. Using the Cox proportional hazards regression model, the risk of developing an NTE in LUSC patients who never became disease-free was significantly higher than that of those who were disease-free (HR = 6.68, 95% CI = 4.25–10.51, false discovery rate [FDR] adjusted q value < 0.05). Similar results were observed in another 21 cancer types (Table 4). Of the remaining 7, we did not see these differences but caution that majority had few patients at risk. We also assessed whether each model satisfied the proportional hazards

Table 5. Comparing Outcomes from the Top Two TSSs that Provided Most Cases (at Least 50) of Each Cancer Type Studied

		OS				PFI				DFI				DSS			
Type	Site	No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]
BLCA	first	52	18	0.58	0.36–0.96*	52	20	0.67	0.42–1.08	25	4	0.61	0.21–1.77	51	12	0.57	0.31–1.05
	second	51	34	1.16	0.79–1.7	51	25	0.92	0.6–1.42	19	2	0.31	0.07–1.32	46	22	1.13	0.7–1.82
	others	308	128			309	125			145	26			300	89		
BRCA	first	150	48	*2.83*	*1.97–4.06***	150	23	*0.93*	*0.59–1.47*	111	3	0.23	0.07–0.74*	140	21	2.13	1.27–3.56**
	second	102	15	*1.08*	*0.62–1.87*	102	16	*0.9*	*0.53–1.53*	83	9	0.8	0.4–1.61	102	10	1.17	0.59–2.3
	others	844	88			844	94			758	72			835	52		
CESC	first	64	29	1.97	1.12–3.47**	64	19	1.35	0.75–2.42	31	8	1.56	0.65–3.75	62	18	1.65	0.85–3.2
	second	54	20	2.27	1.23–4.19**	54	23	2.12	1.23–3.66**	30	4	0.86	0.28–2.62	54	18	2.39	1.24–4.62**
	others	189	22			189	26			115	14			187	18		
COAD	first	173	31	*1.01*	*0.65–1.58*	173	42	*1.02*	*0.69–1.51*	84	8	0.65	0.27–1.61	173	28	*1.52*	*0.89–2.58*
	second	52	11	*0.91*	*0.47–1.76*	52	18	*1.14*	*0.67–1.93*	27	3	0.7	0.2–2.51	52	6	*0.82*	*0.34–1.98*
	others	233	60			233	63			79	13			217	30		
GBM	first	155	128	*1.03*	*0.83–1.27*	155	138	1.21	0.98–1.48	NA	NA	NA	NA	152	124	*1.04*	*0.84–1.3*
	second	93	87	*0.83*	*0.65–1.05*	93	90	0.84	0.66–1.07	NA	NA	NA	NA	90	84	*0.83*	*0.65–1.07*
	others	348	276			348	278			3	2			313	237		
HNSC	first	135	98	1.85	1.37–2.51**	135	59	1.02	0.74–1.42	21	4	0.44	0.12–1.55	121	48	1.6	1.08–2.37*
	second	74	35	1.56	1.05–2.3*	74	25	0.99	0.65–1.51	13	4	1.37	0.46–4.08	73	21	1.37	0.83–2.24
	others	318	90			318	105			99	20			307	61		
KIRC	first	142	39	*1.2*	*0.81–1.79*	141	30	0.85	0.55–1.3	NA	NA	NA	NA	141	22	1.03	0.62–1.73
	second	107	70	*2.72*	*1.94–3.79***	107	59	2.3	1.63–3.26**	29	6	1.14	0.39–3.34	102	45	2.88	1.89–4.38**
	others	288	68			287	71			88	9			282	43		
LGG	first	104	9	*0.65*	*0.33–1.31*	104	18	0.73	0.45–1.21	44	3	0.40	0.12–1.36	104	7	*0.53*	*0.24–1.16*
	second	86	44	*1.73*	*1.17–2.54***	86	59	2.00	1.46–2.75**	NA	NA	NA	NA	81	39	*1.78*	*1.18–2.68***
	others	324	72			324	115			90	17			321	67		
OV	first	111	77	0.59	0.45–0.78**	111	92	0.94	0.74–1.2	63	51	1	0.72–1.41	107	67	0.58	0.44–0.78**
	second	99	72	1.28	0.98–1.68	99	69	1.31	1–1.71*	60	44	1.77	1.24–2.53**	94	67	1.4	1.06–1.87*
	others	372	199			372	252			163	101			344	167		
PRAD	first	97	0	NA	NA	97	13	0.58	0.32–1.04	85	3	0.27	0.08–0.9*	97	0	NA	#N/A
	second	65	0	NA	NA	65	6	0.71	0.33–1.56	43	1	0.26	0.03–1.9	65	0	NA	#N/A
	others	338	10			338	67			212	26			336	5		
SKCM	first	92	55	0.73	0.53–1*	92	73	*0.94*	*0.71–1.23*	NA	NA	NA	NA	89	48	0.71	0.51–1*
	second	68	28	0.54	0.36–0.82**	68	59	*1.51*	*1.12–2.03***	NA	NA	NA	NA	67	24	0.53	0.34–0.83**
	others	295	131			296	176			NA	NA			293	116		

(*Continued on next page*)

Table 5. *Continued*

Type	Site	OS				PFI				DFI				DSS			
		No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]	No.	Event	HR[a]	95% CI[a]
STAD	first	137	56	1.57	1.1–2.24**	137	26	*0.7*	*0.45–1.07*	91	16	*0.9*	*0.48–1.68*	123	17	0.66	0.38–1.15
	second	67	42	1.59	1.08–2.34**	67	38	*1.48*	*1.01–2.19**	29	2	*0.21*	*0.05–0.88**	64	32	1.62	1.04–2.52*
	others	233	71			235	76			137	28			223	52		
THCA	first	93	0	NA	NA	93	1	0.44	0.17–1.12	81	1	0.12	0.02–0.91*	93	0	NA	#N/A
	second	81	0	NA	NA	81	5	0.62	0.24–1.57	61	3	0.65	0.19–2.2	81	0	NA	#N/A
	others	333	16			333	42			216	22			327	7		
UCEC	first	78	10	1.66	0.84–3.26	78	10	1.02	0.54–1.91	70	6	1.16	0.49–2.74	78	3	0.66	0.2–2.12
	second	67	10	0.97	0.5–1.88	67	12	0.95	0.54–1.66	30	5	1.03	0.41–2.59	65	7	0.95	0.43–2.09
	others	402	71			402	91			326	46			402	50		

*, significant model with a nominal $p < 0.05$; **, FDR-adjusted q value < 0.05. *Italic HR and 95% CI* indicate models we recommend for additional testing of PHs assumptions (c.f. Table S1, tab Table5_PHAssumptionTests). K-M plots comparing the survival data from top two TSSs with those from patients of other sites (i.e., not from the top two TSSs) for each cancer type without restricting to TSSs supplying at least 50 cases are shown in Figure S5.

[a]HR and 95% CI were calculated from Cox PHs regression models.

assumptions (Table S1, tab Table4_PHAssumptionTests). Two did not meet the assumptions because of causes that remain to be studied with time-dependent or multivariable models. The K-M plots for these analyses are provided in Figure S4.

The above two study examples, comparing outcome differences between high- and low-stage disease and between patients rendered disease-free or not, demonstrated consistency with our known understanding of the effect of these factors on patient outcomes and, therefore, not only further validate the TCGA-CDR findings but also illustrate how this large clinicopathologic resource can be used to conduct translational cancer research at an unprecedented scale.

TCGA cases were collected from hundreds of sites worldwide, including some tissue bank networks with limited clinical information. There were also differences in the number of retrospectively and prospectively collected samples by site. To address whether clinical data were comparable from site to site, multiple factors, including the completeness of the data, tumor characteristics, and patient characteristics, need to be taken into account. For each disease, we compared the top two TSSs (i.e., the two TSSs that provided the largest number of cases for each disease) against all other submitting sites for the same disease and for each of the four outcome endpoints (Table 5). Additionally, we tested for satisfaction of proportional hazards assumption (Table S1, tab Table5_PHAssumptionTest), and models not satisfying the assumption were flagged. The detailed K-M plots for each of the 33 disease sets are shown in Figure S5.

We observed that, for a highly aggressive tumor like GBM, these top two TSSs (#1 and #2) had similar OS, PFI, and DSS outcomes compared with the disease population from other TSSs, whereas events evaluated by the not-recommended endpoint DFI were too few to be analyzed. For a less aggressive tumor type like BRCA, we recommended use of PFI and DFI without reservation but suggested caution when assessing either OS or DSS. For TSS #1, their clinical data generated worse OS and DSS outcomes and showed no observed difference in their PFI relative to the other sites. However, this same TSS generated a nominally better DFI outcome, having accrued only 3 DFI events. Such nominally inconsistent results suggest that the outcome data from this site need to be further evaluated. TSS #2, on the other hand, consistently generated outcomes comparable with those from other sites across all 4 endpoints.

This simple outcome comparison test suggests that TSS-specific information might need to be taken into account when analyzing the entire body of TCGA clinical data for specific outcomes. Because endpoint-confounding factors from different TSS populations might include patient age, tumor stage/grade, and treatment, TSSs might serve as a proxy for these as well as other unmeasured differences, including incomplete clinical annotation. For this purpose, we have included the key to translate the TSS element of the patient barcode in Table S1, tab TSS_Info.

DISCUSSION

This TCGA-CDR (Table S1, tab TCGA-CDR) was created as the result of a systematic review of TCGA pan-cancer clinical data where we calculated, assimilated, and evaluated four commonly

used clinical outcome endpoints (OS, PFI, DFI, and DSS; Figure 1) for each of the 33 different TCGA cancer types analyzed by the network over the past decade (Table 1). In this effort, we processed and merged data for 10 data elements from the initial patient enrollment data file and subsequent follow-up data files. Another 11 commonly used clinical data elements were also extracted from the initial patient enrollment data file, with quality assessed for inclusion in the TCGA-CDR. A flag for case redaction status was also included; e.g., when the subject withdrew consent or in the cases of genotype mismatch. To accomplish this pan-cancer effort and also address the many data-cleaning issues requiring resolution before finalizing this new scientific resource, we consulted with various TCGA Analysis Working Group (AWG) experts and worked closely with TCGA Biospecimen Core Resource (BCR) personnel who originally collected and curated the data elements for each AWG. Beyond resolving these problematic issues, the quality of each clinical endpoint was independently evaluated to offer recommendations about their use in future studies (Table 3).

Despite the relatively short follow-up time across all TCGA clinical data (Table 2), a limited number of AWG marker papers contained survival analyses employing a few specific endpoints, including GBM, where only OS was used in the initial analysis (Cancer Genome Atlas Research Network, 2008), OV (Cancer Genome Atlas Research Network, 2011) and LGG (Cancer Genome Atlas Research Network et al., 2015), where OS was also used, and UCEC, where PFI was used (Cancer Genome Atlas Research Network et al., 2013). For future GBM studies, we can now recommend the use of OS, PFI, and DSS (as an approximation) with confidence, and for UCEC and OV studies, we can now confidently recommend the use of all four clinical endpoints, including DSS, as an approximation (Table 3). In several TCGA studies, survival analyses were not reported, for example in prostate adenocarcinoma (PRAD) and BRCA (Cancer Genome Atlas Network, 2012; Cancer Genome Atlas Research Network, 2015). For both PRAD and BRCA, we can now recommend use of PFI and DFI but advise caution when interpreting OS or DSS for BRCA and do not recommend using OS or DSS for PRAD (Table 3). Of note, DFI was effectively used in more recent BRCA racial disparity studies (Huo et al., 2017; Keenan et al., 2015).

Although OS is the most accurately derived endpoint from the TCGA clinical data as curated in the TCGA-CDR, our assessment shows that OS is an appropriate endpoint for many but not all cancer types. For aggressive TCGA cancer types like GBM, where the median OS event time was 12.6 months, and the median follow-up of 12 months allowed capturing events in 82.4% of the cases, OS is an appropriate clinical endpoint. Likewise, OV cases in the TCGA cohort with their median OS event time of 35.3 months and an event rate of 59.5% (349 of 587 cases) proved to be sufficiently aggressive to support this as an appropriate clinical endpoint (Cancer Genome Atlas Research Network, 2011). However, for less aggressive cancer types like BRCA, appropriate use of OS depends on BRCA subtype. There are 4 primary intrinsic subtypes of BRCA (Perou et al., 2000), including the most aggressive basal-like subtype, which commonly recurs within a few years, in contrast to the least aggressive Luminal A subtype, which may require 10 years or more to recur. Hence, given a relatively short follow-up time, OS may be a suitable endpoint for the basal-like subtype but not for the Luminal A subtype. For an even less aggressive TCGA cancer type like PRAD, where there were only 10 OS events out 500 cases, OS is clearly not a suitable study endpoint (Table 3).

In recommending survival endpoint choices within the TCGA-CDR, our analysis emphasizes strengths and limitations behind each of the four different endpoint calculations. For OS, events are clearly defined by the time of patient death. However, for much of the TCGA clinical data, DSS had to be approximated because of lack of absolute verification of the cause of death. Given the relatively short clinical follow-up records for most of the TCGA cohorts, PFI and DFI might generally be considered better clinical endpoint choices than OS and DSS because patients normally develop disease recurrence before death and, therefore, more endpoint events are recorded during the follow-up period. PFI in particular was derived with high confidence and is a recommended clinical endpoint choice for 27 of the 33 pan-cancer types. In five others, its use is recommended with caution, and for one (LAML), outcome data are lacking and PFI is not available. Unlike PFI, calculation of DFI requires that patients be documented as free of disease at a specific point in time following their initial diagnosis; because this explicit point in time was not available in the TCGA clinical dataset, here DFI event time (or censoring) had to be calculated from the day of initial diagnosis for cases where there was evidence that primary treatment (e.g., surgical excision) rendered the patient disease-free. In analysis (Table 4), we used 90 days past diagnosis as an approximate duration of the common pan-cancer primary treatment interval. More precise estimates of treatment duration may be available per cancer type. Among the 4 endpoints, PFI is generally considered a more informative endpoint for TCGA pan-cancer studies; however, DFI is also an excellent endpoint and should be considered suitable for further research into many TCGA cancer types.

For clinical survival endpoint analyses, all clinical data we have processed can be used as recommended. But for integrative analysis with molecular/genomic data, caution is warranted in two respects. First, our recommendations are based on baseline survival models. The inclusion of molecular subgroups as predictors begins to partition the sample sets, potentially compromising the statistical significance of apparent outcome differences (Peduzzi et al., 1995). Therefore, conclusions drawn from cross-correlating TCGA molecular data or tumor subgroups with TCGA-CDR outcome data should be further validated on independent tumor datasets. Second, in general, we recommend use of only molecular data from primary tumors because the matching clinical data, including important temporal information, were collected relatively completely for patients at the time of initial diagnosis. In particular, skin cutaneous melanoma (SKCM) is unique among the TCGA tumor types because, among 470 tumors, only 103 were primary tumors, whereas 296 were regional lymph node metastases, and 68 were distant metastases. This is in stark contrast to other TCGA cancer types where few metastatic tumors were collected. Very few SKCM metastatic tumors had a matching primary tumor, in contrast to other TCGA cancer types where the few metastatic tumors

all had matching primary samples. Thus, for SKCM outcome correlations, we recommend using only the limited number of primary cases, although the SKCM nodal metastases for stage III cases could be studied as a discrete group.

The curated clinical endpoint results as presented in the TCGA-CDR are consistent with independent outcome reports from other comparable cancer cohorts. For example, the OS and PFI event times we derived for GBM are consistent with those in the literature (Stupp et al., 2005). We compared the TCGA-CDR-determined breast cancer outcomes with well-established survival differences reported between patients with ER+ and ER− tumors (Ren et al., 2014; Saphner et al., 1996; Yu et al., 2012). As expected, patients with ER+ tumors had significantly better outcome differences using PFI, DFI, and DSS as the end points.

Data from the TCGA-CDR not only enable investigators to generate results consistent with those from independent studies but also invite exploration into biologically relevant questions across multiple cancer types at an unprecedented scale. For example, we demonstrate that TCGA patients who were never disease-free after initial diagnosis show significantly higher odds of developing an NTE compared with those who were once disease-free after treatment (Table 4). This finding also highlights the importance of selecting the most appropriate outcome variables (Anderson et al., 1983, 2008). Although it may not be an intuitively surprising observation that patients never made free of disease are likelier to develop early tumor progression and die than those rendered free of disease after diagnosis (Bachy et al., 2010; Schnitt, 2003), the sheer volume of available TCGA clinical data enabled us to statistically confirm this expectation with sufficient power even when conditioning on a landmark event such as end of treatment (here approximated at 3 months).

The relationship of clinical endpoints that may act as surrogates for overall survival are of great importance to therapeutic trials (Chibaudel et al., 2011; Oxnard et al., 2012; Shi and Sargent, 2009). Differences in interpretation (e.g., disease-specific cause of death, evidence of progression) and accuracy of measurement (e.g., time of event versus time of detection) influence this decision (Blumenthal et al., 2015; Johnson et al., 2003). This is particularly true for cancers like BRCA and PRAD, in which we see that OS may not be an appropriate endpoint without sufficient follow-up time. When we compared survival for patients with high-stage versus low-stage disease, we confirmed a significantly worse outcome for those with high-stage disease (Figure S3; Table S1, tab Figure 2EFG_AdditionalInfo) and demonstrated that their logHRs measured for OS, PFI, and DSS are strongly and significantly correlated (Figures 2E–2G). This observation validates the common practice of clinical trials to choose intermediate endpoints like disease progression or recurrence events when comparing interventions for less aggressive cancers, where overall or disease-specific survival outcomes might otherwise require decades of follow-up observations.

Previous TCGA studies have reported important clinical findings of translational significance on interim smaller cohorts. For example, in the OV marker paper (Cancer Genome Atlas Research Network, 2011), OS data were used to generate a 193-gene transcriptional signature for survival prediction, and its predictive power was validated using several other independent datasets. In the LGG marker paper (Cancer Genome Atlas Research Network, 2015), analysis of OS showed that patients diagnosed with an *IDH1* and *IDH2* (two very similar genes, hereafter referred to collectively as *IDH*) mutation with or without 1p/19q codeletion had substantially longer OS than did patients who had wild-type IDH, proving that IDH-1p/19q status represents a more robust survival predictor than LGG histologic subtype. Indeed, the clinical and molecular work of this paper factored into evidence used by the World Health Organization (WHO) to support their 2016 diagnostic update for glioma (Louis et al., 2016). Importantly, in this TCGA-CDR, the number of cases for OV, LGG, and others have significantly increased and now have longer follow-up times, which provides greater statistical power for future outcome analyses of these or other TCGA cancer types.

As for all clinical datasets, using TCGA data for outcome analyses requires picking relevant endpoints, determining appropriate statistical methods, and carefully validating models. In providing this newly curated TCGA-CDR database, we have considered and wish to point out three important issues that should be noted by future users of this resource: potential confounding factors, competing outcome risks (Fine and Gray, 1999), and model assumptions.

Confounding Factors

When confounding factors are present but excluded from the model, bias can manifest as either over- or underestimation of a true effect. For example, in breast cancer racial disparity studies, major gene expression differences have been observed between black and white patients, but, after adjusting for molecular subtypes, such differences can be significantly reduced or nullified (Huo et al., 2017; Keenan et al., 2015). Treatment effects are also potential confounders (see below) that should be considered when available and adjusted for appropriately. In certain cases, a proxy for standard of care, such as age, treating hospital, and year of diagnosis, can alleviate some of the bias when treatment is unknown. For modeling decisions in this regard, we encourage the use of the reporting recommendations for tumor marker prognostic studies (REMARK) (Altman et al., 2012; McShane et al., 2005). Similar to Consolidated Standards of Reporting Trials (CONSORT) guidelines for reporting randomized clinical trials (Schulz et al., 2010), REMARK seeks to improve the transparency and reproducibility of prognostic marker studies.

Competing Outcome Risks

In our determination of DSS, DFI, and PFI endpoints, patients who died without experiencing the event of interest and were also disease-free were censored. In this way, we assumed that if a patient had not died of other causes, then (s)he might have eventually died of the index cancer. However, there may be situations in which this assumption is not desirable, such as if we were to want to estimate the effect of a predictor, say treatment, on the risk of non-index cancer death, say development of secondary cancer or cardiovascular disease. We looked for this potential confounding effect of non-index cancer death in our application example of cancer stages (Figures 2E–2G) and compared results across cancer types of three different endpoints but concluded that, for this example, the effect of competing risk

was minimal (STAR Methods). However, to enable modeling of competing risk by future users of this resource, we derived additional endpoints to indicate non-index death (Table S1, tab ExtraEndpoints).

Model Assumptions

When applying the Cox proportional hazards (PH) model in particular, the PH assumption must be examined (Grambsch and Therneau, 1994). In the disease and outcome examples we discussed, we emphasize that most of our models satisfied the Cox PH assumption with only few exceptions, but these exceptions deserve further exploration to try to discern why (e.g., aging effects, decreasing influence over time) they violated the PH assumption so that more accurate estimates of the HRs can be made.

Despite thorough efforts to clean the data and resolve issues, there remain important use limitations with these curated data that must be appreciated by all who access this TCGA clinical data compendium. First and foremost, because TCGA was designed primarily for molecular studies, clinical data collection was secondary, and reporting for a number of data fields was not required by the program. Initial case selection criteria were for untreated primary cases with appropriately banked tissues from multiple institutions, and such cases, thus, do not constitute a consecutive series. In TCGA, TSSs were required to provide initial clinical information as samples were accrued and 1 year later additional follow-up clinical information where possible. For prospectively collected samples, the follow-up time could be as short as only 1 year. In addition, the follow-up data were not collected uniformly for each of the different tumor types/studies. Although there were technical difficulties that prevented some sites from supplying follow-up clinical data, a number of AWGs were very proactive in working with TSSs and the BCR to ensure that this follow-up information was as accurate and up-to-date as possible. Having participants from the TSSs in these AWGs greatly improved the overall quality of the resulting clinical data for these TCGA cancer types. Some rules for clinical data collection had to be changed over time, which was unavoidable for a multi-national project that improved its execution over its 10-year time frame. With additional follow-up data collection, the value of this TCGA clinical dataset would grow. Although TCGA funding has ended, we encourage the development of coordinated efforts to follow up on patients who were alive at last follow-up.

Following from the above, TCGA-CDR does not contain cancer treatment history (for the reasons we provide in the STAR Methods). Some treatment information is available for 32 of the 33 TCGA cancer types, although this was not TCGA-required data annotation for samples accruing from each TSS, so not all cases will be annotated. To complete treatment annotations on more 11,000 TCGA cancer cases representing 33 different cancer types and then to assess the completeness, accuracy, and outcome effect of these annotated treatment data would require a major undertaking that is beyond the scope and means of this current resource curation. There are different therapies for different cancer types and subtypes, and even within one cancer type or subtype there are multiple treatment regimens. For these reasons, we feel that analyzing therapy within the context of a particular tumor type (and/or subtype) may be more appropriate than pan-cancer generalizations. When patients were treated with specific therapies, the benefits of such treatments can be effectively assessed by endpoints such as DFI and PFI.

A second use limitation of this curated TCGA-CDR is that TCGA samples were accrued both retrospectively and prospectively, and patients were clinically followed up according to local clinic schedules that might be disease- and site-specific for the recording of disease recurrence and patient vital status; thus, there was no TCGA-specified clinical follow-up plan, given the program's primary emphasis on tumor molecular characterization. Last, almost all TCGA-acquired tumor samples and, therefore, the genomic and molecular data derived from them, come from single sections of primary tumors in newly diagnosed patients; the resulting genomic and molecular data do not explicitly capture any spatial or temporal aspects of tumor heterogeneity that could potentially represent another patient outcome variable. This limitation, though, is not unique to TCGA but true to any static primary tumor study.

In summary, this work represents the first ever comprehensive effort to systematically process TCGA pan-cancer clinical data. We assembled and integrated all of the acquired clinical data files, reviewed and carefully analyzed dozens of different data elements important to cancer research, resolved over 1,000 quality assessed (QA) issues, and generated four commonly used clinical outcome endpoints for each of the 11,160 tumor cases: OS, PFI, DFI, and DSS. Using well established and newly developed analysis methods for each tumor type, we quality-scored all four outcome determinations and further provided tumor-specific recommendations for their use in future studies. We also show that the resulting TCGA-CDR yields outcome endpoints consistent with independent non-TCGA study findings for different tumor types and demonstrate how this resource offers new opportunities to produce biologically insightful observations at unprecedented clinical scale. In recognizing the limitations inherent within the TCGA-CDR and providing critical guidance and recommendations for its appropriate use, it has become abundantly clear that future large-scale molecular studies of human diseases must also systematically collect clinicopathologic, treatment, and outcome event data adhering to the highest clinical research standards. Its limitations notwithstanding, TCGA-CDR presents a standardized dataset with a transparent derivation of four clinical outcome endpoints and resolution of quality concerns, enabling translational studies at both pan-cancer and individual disease levels. Adoption of this dataset by future studies will improve comparability of the results between studies to afford better interpretation and support reproducibility.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - The TCGA clinical data of 11,160 patients across 33 tumor types were analyzed.

- METHOD DETAILS
  - Clinical data
  - Handling of special cases and problems in clinical data files
  - Choice of time zero for time-to-event calculations
  - Definitions and derivation of clinical survival outcome endpoints
  - Assessment of clinical endpoints of OS, PFI, DFI, DSS
  - Competing risk
  - Other clinical data fields processed
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, and four tables and can be found with this article online at https://doi.org/10.1016/j.cell.2018.02.052.

ACKNOWLEDGMENTS

We thank the patients who participated in this study. We are grateful for the clinical staff of TCGA Tissue Source Sites and the staff of the TCGA Biospecimen Core Resource for collecting, compiling, and processing the clinical data for TCGA that laid the foundation for this study. We thank Ms. Anupama Praveen-Kumar and Drs. Joseph Vockley and Praveen-Kumar Raj-Kumar for helpful discussions and Mr. Peter T. Hu for creating the graphical abstract. Members of the TCGA PanCanAtlas Network and especially those from the Immune Response AWG participated in discussions as the project progressed. The study was supported by W81XWH-12-2-0050, HU0001-16-2-0004 from the U.S. Department of Defense through the Henry M. Jackson Foundation for the Advancement of Military Medicine. The views expressed in this article are those of the authors and do not reflect the official policy of the department of Army/Navy/Air Force, Department of Defense, or U.S. government. The study was also supported by the TCGA grants U54 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, P30 CA016672.

AUTHOR CONTRIBUTIONS

Conceptualization, H.H.; Methodology, J.L. and L.M.P.; Formal Analysis, J.L., L.M.P., D.A.L., and L.O.; Investigation, T.L., K.A.H., and A.D.C.; Resources, T.L., K.A.H., A.D.C., and TCGA Network; Data Curation, J.L., T.L., K.A.H., and TCGA Network; Writing – Original Draft, H.H., J.L., K.A.H., T.L., and A.J.K., Writing – Review & Editing, H.H., J.L., C.C.B., L.M.P., K.A.H., A.J.L., A.V.L., D.A.L., D.M.W., V.T., A.J.K., and C.D.S.; Visualization, J.L.; Funding Acquisition, C.D.S. and H.H.; Supervision and Project Administration, H.H.

DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine, Inc. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine, Inc. Andrew D. Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of Astrazeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck, Inc. Kyle R. Covington is an employee of Castle Biosciences, Inc. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine, Inc. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine, Inc. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck, Inc. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO, Inc. Bin Feng is an employee and shareholder of TESARO, Inc. Carter Van Waes received research funding for the study of IAP inhibitor ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges, Inc. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth, Inc. and a shareholder and advisory board member of Tempus, Inc. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invitae.

Received: July 21, 2017
Revised: November 11, 2017
Accepted: February 20, 2018
Published: April 5, 2018

REFERENCES

Altman, D.G., McShane, L.M., Sauerbrei, W., and Taube, S.E. (2012). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration. PLoS Med. *9*, e1001216.

Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., and Winchester, D.P. (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA Cancer J. Clin. *67*, 93–99.

Anderson, J.R., Cain, K.C., and Gelber, R.D. (1983). Analysis of survival by tumor response. J. Clin. Oncol. *1*, 710–719.

Anderson, J.R., Cain, K.C., and Gelber, R.D. (2008). Analysis of survival by tumor response and other comparisons of time-to-event by outcome variables. J. Clin. Oncol. *26*, 3913–3915.

Bachy, E., Brice, P., Delarue, R., Brousse, N., Haioun, C., Le Gouill, S., Delmer, A., Bordessoule, D., Tilly, H., Corront, B., et al. (2010). Long-term follow-up of patients with newly diagnosed follicular lymphoma in the prerituximab era: effect of response quality on survival–A study from the groupe d'etude des lymphomes de l'adulte. J. Clin. Oncol. *28*, 822–829.

Blumenthal, G.M., Karuri, S.W., Zhang, H., Zhang, L., Khozin, S., Kazandjian, D., Tang, S., Sridhara, R., Keegan, P., and Pazdur, R. (2015). Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non-small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. J. Clin. Oncol. *33*, 1008–1014.

Cancer Genome Atlas Research Network, Brat, D.J., Verhaak, R.G., Aldape, K.D., Yung, W.K., Salama, S.R., Cooper, L.A., Rheinbay, E., Miller, C.R., Vitucci, M., Morozova, O., et al. (2015). Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N. Engl. J. Med. *372*, 2481–2498.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Cancer Genome Atlas Network (2015). Genomic Classification of Cutaneous Melanoma. Cell *161*, 1681–1696.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*, 1061–1068.

Cancer Genome Atlas Research Network (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*, 519–525.

Cancer Genome Atlas Research Network, Kandoth, C., Schultz, N., Cherniack, A.D., Akbani, R., Liu, Y., Shen, H., Robertson, A.G., Pashtan, I., Shen, R., et al. (2013). Integrated genomic characterization of endometrial carcinoma. Nature *497*, 67–73.

Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of urothelial bladder carcinoma. Nature *507*, 315–322.

Cancer Genome Atlas Research Network (2015). The Molecular Taxonomy of Primary Prostate Cancer. Cell *163*, 1011–1025.

Cancer Genome Atlas Research Network; Albert Einstein College of Medicine; Analytical Biological Services; Barretos Cancer Hospital; Baylor College of Medicine; Beckman Research Institute of City of Hope; Buck Institute for Research on Aging; Canada's Michael Smith Genome Sciences Centre; Harvard Medical School; Helen F. Graham Cancer Center &Research Institute at Christiana Care Health Services; HudsonAlpha Institute for Biotechnology; ILSbio, LLC; Indiana University School of Medicine; Institute of Human Virology; Institute for Systems Biology; International Genomics Consortium; Leidos Biomedical; Massachusetts General Hospital; McDonnell Genome Institute at Washington University; Medical College of Wisconsin; Medical University of South Carolina; Memorial Sloan Kettering Cancer Center; Montefiore Medical Center; NantOmics; National Cancer Institute; National Hospital, Abuja, Nigeria; National Human Genome Research Institute; National Institute of Environmental Health Sciences; National Institute on Deafness &Other Communication Disorders; Ontario Tumour Bank, London Health Sciences Centre; Ontario Tumour Bank, Ontario Institute for Cancer Research; Ontario Tumour Bank, The Ottawa Hospital; Oregon Health &Science University; Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center; SRA International; St Joseph's Candler Health System; Eli &Edythe L. Broad Institute of Massachusetts Institute of Technology &Harvard University; Research Institute at Nationwide Children's Hospital; Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University; University of Bergen; University of Texas MD Anderson Cancer Center; University of Abuja Teaching Hospital; University of Alabama at Birmingham; University of California, Irvine; University of California Santa Cruz; University of Kansas Medical Center; University of Lausanne; University of New Mexico Health Sciences Center; University of North Carolina at Chapel Hill; University of Oklahoma Health Sciences Center; University of Pittsburgh; University of São Paulo, Ribeir ão Preto Medical School; University of Southern California; University of Washington; University of Wisconsin School of Medicine &Public Health; Van Andel Research Institute; Washington University in St Louis (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378–384.

Chibaudel, B., Bonnetain, F., Shi, Q., Buyse, M., Tournigand, C., Sargent, D.J., Allegra, C.J., Goldberg, R.M., and de Gramont, A. (2011). Alternative end points to evaluate a therapeutic strategy in advanced colorectal cancer: evaluation of progression-free survival, duration of disease control, and time to failure of strategy–an Aide et Recherche en Cancerologie Digestive Group Study. J. Clin. Oncol. *29*, 4199–4204.

Clark, T.G., Bradburn, M.J., Love, S.B., and Altman, D.G. (2003). Survival analysis part I: basic concepts and first analyses. Br. J. Cancer *89*, 232–238.

Fine, J.P., and Gray, R.J. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. J. Am. Stat. Assoc. *94*, 496–509.

Giobbie-Hurder, A., Gelber, R.D., and Regan, M.M. (2013). Challenges of guarantee-time bias. J. Clin. Oncol. *31*, 2963–2969.

Grambsch, P.M., and Therneau, T.M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika *81*, 515–526.

Harrell, F.E., Jr., Lee, K.L., and Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. *15*, 361–387.

Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., et al.; Cancer Genome Atlas Research Network (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell *158*, 929–944.

Hudis, C.A., Barlow, W.E., Costantino, J.P., Gray, R.J., Pritchard, K.I., Chapman, J.A., Sparano, J.A., Hunsberger, S., Enos, R.A., Gelber, R.D., and Zujewski, J.A. (2007). Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J. Clin. Oncol. *25*, 2127–2132.

Huo, D., Hu, H., Rhie, S.K., Gamazon, E.R., Cherniack, A.D., Liu, J., Yoshimatsu, T.F., Pitt, J.J., Hoadley, K.A., Troester, M., et al. (2017). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol. *3*, 1654–1662.

Johnson, J.R., Williams, G., and Pazdur, R. (2003). End points and United States Food and Drug Administration approval of oncology drugs. J. Clin. Oncol. *21*, 1404–1411.

Keenan, T., Moy, B., Mroz, E.A., Ross, K., Niemierko, A., Rocco, J.W., Isakoff, S., Ellisen, L.W., and Bardia, A. (2015). Comparison of the Genomic Landscape Between Primary Breast Cancer in African American Versus White Women and the Association of Racial Differences With Tumor Recurrence. J. Clin. Oncol. *33*, 3621–3627.

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. *131*, 803–820.

Maller, R.A., and Zhou, S. (1994). Testing for Sufficient Follow-Up and Outliers in Survival Data. J. Am. Stat. Assoc. *89*, 1499–1506.

McShane, L.M., Altman, D.G., Sauerbrei, W., Taube, S.E., Gion, M., and Clark, G.M.; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics (2005). REporting recommendations for tumour MARKer prognostic studies (REMARK). Br. J. Cancer *93*, 387–391.

Oxnard, G.R., Morris, M.J., Hodi, F.S., Baker, L.H., Kris, M.G., Venook, A.P., and Schwartz, L.H. (2012). When progressive disease does not mean treatment failure: reconsidering the criteria for progression. J. Natl. Cancer Inst. *104*, 1534–1541.

Peduzzi, P., Concato, J., Feinstein, A.R., and Holford, T.R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J. Clin. Epidemiol. *48*, 1503–1510.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R., and Feinstein, A.R. (1996). A simulation study of the number of events per variable in logistic regression analysis. J. Clin. Epidemiol. *49*, 1373–1379.

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. Nature *406*, 747–752.

Punt, C.J., Buyse, M., Köhne, C.H., Hohenberger, P., Labianca, R., Schmoll, H.J., Påhlman, L., Sobrero, A., and Douillard, J.Y. (2007). Endpoints in adjuvant treatment trials: a systematic review of the literature in colon cancer and proposed definitions for future trials. J. Natl. Cancer Inst. *99*, 998–1003.

Ren, Y., Black, D.M., Mittendorf, E.A., Liu, P., Li, X., Du, X.L., He, J., Yang, J., Hunt, K.K., and Yi, M. (2014). Crossover effects of estrogen receptor status on breast cancer-specific hazard rates by age and race. PLoS ONE *9*, e110281.

Saphner, T., Tormey, D.C., and Gray, R. (1996). Annual hazard rates of recurrence for breast cancer after primary therapy. J. Clin. Oncol. *14*, 2738–2746.

Schnitt, S.J. (2003). Risk factors for local recurrence in patients with invasive breast cancer and negative surgical margins of excision. Where are we and where are we going? Am. J. Clin. Pathol. *120*, 485–488.

Schulz, K.F., Altman, D.G., and Moher, D.; CONSORT Group (2010). CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Med. *8*, 18.

Shen, P. (2000). Testing for sufficient follow-up in survival data. Stat. Probab. Lett. *49*, 313–322.

Shi, Q., and Sargent, D.J. (2009). Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. Int. J. Clin. Oncol. *14*, 102–111.

Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., et al.; European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups; National Cancer Institute of Canada Clinical Trials Group (2005). Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N. Engl. J. Med. *352*, 987–996.

Tai, P., Yu, E., Cserni, G., Vlastos, G., Royce, M., Kunkler, I., and Vinh-Hung, V. (2005). Minimum follow-up time required for the estimation of statistical cure of cancer patients: verification using data from 42 cancer sites in the SEER database. BMC Cancer *5*, 48.

Verhaak, R.G., Hoadley, K.A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M.D., Miller, C.R., Ding, L., Golub, T., Mesirov, J.P., et al.; Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell *17*, 98–110.

Yu, K.D., Wu, J., Shen, Z.Z., and Shao, Z.M. (2012). Hazard of breast cancer-specific mortality among women with estrogen receptor-positive breast cancer after five years from diagnosis: implication for extended endocrine therapy. J. Clin. Endocrinol. Metab. *97*, E2201–E2209.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
TCGA clinical data (33 tumor types; 225 files)	Genomic Data Commons data portal	https://gdc-portal.nci.nih.gov/legacy-archive/
Tissue Source Site Codes	National Cancer Institute Genomic Data Commons	https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tissue-source-site-codes
Sample Submission Form	National Cancer Institute GDC Legacy Archive	https://portal.gdc.cancer.gov/legacy-archive/search/f
TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR)	This paper	Table S1
Software and Algorithms		
R 3.2.2	**R** Foundation for Statistical Computing	https://www.r-project.org/about.html

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact Hai Hu (h.hu@wriwindber.org).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The TCGA clinical data of 11,160 patients across 33 tumor types were analyzed.
Commonly used clinical data elements were quality assessed (QA), integrated and processed with the help from disease experts and TCGA Biospecimen Core Resource. Clinical outcome endpoints of overall survival (OS), progression-free interval (PFI), disease-free interval (DFI), and disease-specific survival (DSS) were derived. Each qualifying end point was subjected to multi-step assessment, and led to varying endpoint recommendations per tumor type. Validation and Application of the TCGA CDR data were performed by examples.

METHOD DETAILS

Clinical data

The TCGA clinical data were downloaded from the data portal of Genomic Data Commons (GDC, https://gdc-portal.nci.nih.gov/legacy-archive/), by selecting ''Biotab'' as the Data Format, ''Clinical Data'' as the Data Type, and ''Clinical'' as the Platform. From the total of 225 files of TCGA clinical data, 130 initial enrollment and follow-up files were used in this paper. Other files for radiation or pharmaceutical treatments were not used. Note that all TCGA molecular data are also available from GDC. The same TCGA barcode structure is used for both clinical data and molecular data, with the first portions of the barcode standing for TSS, patient, and sample followed by tissue aliquots and other information important for molecular studies. This barcode structure enables integration of patient-based clinical data with sample-based molecular data.

TCGA network collected tumors from 161 tissue source sites (TSSs) across the world, acquiring tumors from 11,160 patients of 33 different cancer types (see Abbreviations); these tumors were originally diagnosed from 1978 to 2013 with the median diagnosis year of 2009. With the exception of SKCM (see Results), it was dominantly primary tumors that were included in this pan-cancer collection. Clinical data collection started in 2007. Each TSS completed the initial enrollment data form when a case passed pathology and molecular qualification metrics at the BCR. Follow-up data was then provided one year or later after specimen submission; thus, the follow-up date is generally unique to each case. A prospectively constituted Disease Working Group (DWG) represented each TCGA cancer type and determined the specific clinical data fields to be collected for study by each AWG. While these working groups could add unlimited questions to the data forms, they were limited in the number of fields that would be contractually required by the submitting sites, which was usually around 5-10 fields. For example, the breast cancer DWG prioritized ER, PR and HER2 status over grade for collection. During this clinical data collection process, the TCGA program made a handful of site visits to TSSs to ensure that data collection was being completed correctly. During these visits, source documents were reviewed and compared to the information submitted to TCGA. The BCR also spoke very frequently with TSSs during the years when they were submitting clinical data. The BCR had internal applications that validated and tracked the data following established QA rules that are based on logical relationships between data elements. Data fields and individual data entries were later reviewed by each AWG for

acceptance or censorship from use in the primary TCGA marker paper. Due to these individualized DWG and AWG processes, the datasets for each of the 33 TCGA studies are unique but there are shared data elements common to all cancer types. Across all TCGA studies, hundreds of data elements were included in the initial enrollment forms and subsequent follow-up forms.

In this study we did not analyze cancer treatment data but we would like to provide some information about TCGA treatment history data status here. Treatment data are available for 32 of the 33 cancer types although not for all the cases, because they were not required data fields when the data were collected. We did not include these data in this TCGA-CDR as we understand that the treatment history may not be complete and may not show an accurate overview of treatment for the following reasons known to us: (1) Not all TSSs were where the treatment was performed thus treatment history for patients in those TSSs may not be available; (2) Prospectively enrolled patients only had updates 1 year after sample collection so treatment information would likely be incomplete; (3) There is tremendous heterogeneity within treatment data for patients accepted for TCGA studies from all around the world. We believe that it will be a major undertaking to assess the completeness, accuracy, and value of the treatment data, which demands the establishment of a different AWG to include oncologists and additional disease domain experts. For example, there are different therapies for different cancer types and subtypes, and even within one cancer type or subtype there are multiple treatment regimens. For these reasons we feel that analyzing therapy within the context of a particular tumor type (and/or subtype) may be more appropriate than pan-cancer generalizations. When patients were treated with specific therapies the benefits of such treatments can be effectively assessed by endpoints such as DFI and PFI.

Handling of special cases and problems in clinical data files

In processing the clinical data, we encountered over 1000 cases with apparent or real problems.

1) There were 7 Stage 0 cases which seemed to be an error since TCGA studies were focusing on invasive cancers but Stage 0 cancers are *in situ.* However all of these Stage 0 cases were from the SKCM study, where patients' clinical data were referring to the time of diagnosis. For all these Stage 0 cases, it was the subsequently developed distant metastasis tumors that were used in molecular profiling. Of the 470 tumors used in the SKCM study, only 103 were primary tumors, and 296 were regional lymph node metastases. The remaining 68 were distant metastases, and those metastatic tumors did not have a matching primary tumor.
2) There were 483 cases of patients who died "With Tumor" but without a defined "New NTE," and they were not *de novo* Stage IV. This left us unsure whether these patients really died of tumor, and if so were the new tumors new primary tumors or recurrent tumors of any type. In a previous breast cancer study (Huo et al., 2017) there were 17 such cases, and the authors decided to exclude them from the breast cancer-free interval endpoint analysis (equivalent to DFI here) after reviewing each of them by checking into the original clinical database at BCR. For the current study, after a careful vetting we decided to also exclude these 483 patients from DFI and but included them in OS, PFI and DSS.
3) There were 62 patients who were "Dead" with "Tumor Free," but there was a defined "New NTE." This apparent data inconsistency would affect derivation of DSS but not the other 3 endpoints. This apparent data inconsistency might be what really had happened, as the patient might have developed a new NTE but then treated again to enter a "Tumor Free" status, and subsequently died. In this scenario the case should be censored. Equally, there could be an error in vital status or tumor status, yet we have no evidence which scenario is true. Thus we considered all 62 cases as censored for DSS. In contrast, the DFI endpoint is defined at the time of the NTE.
4) There were10 patients who were "Dead" with "Tumor Free," but the exact cancer type was given as "Cause of Death." Sieving through the BCR records, we found that three patients had no residual tumor (R0) after surgery, but died 66, 113, and 436 days respectively after diagnosis. Two patients had an "unknown" tumor status in the xml file, but there were notions that they had progressive diseases. One other patient had additional update 2 years later with tumor status "unknown" but the information was not updated in the final xml file. For the remaining 4 patients, there was no explanation found. Combining these additional lines of evidence with the caution we provided before regarding the field of "Tumor Status," we resolved these conflicts based on the field of "Cause of Death," with a caveat that clinically sometimes it is difficult to pinpoint specifically what causes someone to die and or if a symptom or illness is related to the cancer.
5) There were 797 of the 3346 "New NTE" that did not have a new tumor type specified. In the strict definition of PFI, the progression should be referring to the progression of the initial primary tumor and thus these cases would have been excluded. However, during data analysis the PanCanAtlas Pan-Immune AWG decided to adopt a more relaxed definition of PFI to include any new tumor as an event so these 797 were included in PFI. We assessed that the rate of a positive response in the "new primary tumor in other organ" field among all 797 new NTEs is very small, thus these 797 cases can be included in PFI analysis to minimize selection bias.
6) There were 6 cases showing a negative value ranging from −1 to −64 in the field "last_contact_days_to," suggesting that the patient was last contacted before diagnosis. There were also 6 cases with an AJCC Stage IV showing a negative value ranging from −4 to −359 in "new_tumor_event_dx_days_to," suggesting that their new NTEs were reported before the patient was diagnosed with cancer. We carefully reviewed the original data, and found different reasons for those negative values but concluded that the solution to those cases were to set the value of these fields to "0."

7) There were 46 patients whose ages were 90, and who have equal values in ''birth_days_to.'' This artificial set of ages is because of the de-identification requirement specified in Health Insurance Portability and Accountability Act of 1996. Users need to be cautious that patients shown at an age of 90 are actually 90 years of age or older.
8) Priority of data files: When there is inconstancy between follow-up time from files of different recording dates, should the data from a later follow-up file take the priority? The answer is no. The recording dates for follow-up forms are administrative only and may not correlate with patient history. Therefore, we decided that the record with a longer follow-up time would be preferred when determining event times.
9) There was one patient whose vital status in the enrollment data file was ''Alive'' with a days_to_last_followup, yet in a follow-up file the vital status was ''Dead'' but without ''days_to_death.'' We choose to use the data in the enrollment data file since the timing of the subsequent death could not be determined.
10) There was one patient with OV showing a tumor with grade 4. Grade 4 is not a valid value for OV. However since this is the final data in the TCGA dataset and we have no evidence to support re-grading of the tumor, we left it as grade 4.

Choice of time zero for time-to-event calculations

In TCGA-CDR, we chose the date of diagnosis as time zero for time-to-event calculations for the following reasons. First, TCGA focused on untreated primary tumors. Of all the 33 cancer types, only SKCM contained a significant portion of the samples that represented either local or distant metastatic tumors for which there were no matching primary tumors. All other cancer types had no more than a few metastatic tumors which were all accompanied with matching primary tumors. There were a few tumors identified after submission as having prior treatment, most were for prior or other malignancies. Based on treatment type TCGA determined if the case was acceptable or unacceptable for use in molecular studies. Only 42 (0.3%) patients with neoadjuvant treatment made it into TCGA and were marked with a notification by the TCGA program office.

For any ''primary'' tumors, the tissue sample was removed from the patient at or close to the time of diagnosis. A field of ''days_to_sample_procurement'' is available in the tumor-sample data file for each disease by going to the link of https://portal.gdc.cancer.gov/legacy-archive/search/f to search in the ''File'' tab using a disease-specific file name, for example for BRCA the search term will be ''nationwidechildrens.org_ssf_tumor_samples_brca.txt.'' For any other disease the abbreviated disease name will be used in the place of BRCA in the search term. For metastatic tumors, the values in the field of ''days_to_sample_procurement'' are relatively large but for primary tumors, the values in this field are small. For example for BRCA, the mean was 34 days, and the median was 21 days. Within such a small number of days we do not feel that tumors would have developed so much as to impact the clinical outcomes, and that using date of diagnosis as time zero would be a good approximation for the date of sample procurement.

Finally, in TCGA data collection, dates were requested which were later processed to derive temporal reference to time zero for compliance with HIPAA regulations. The date of diagnosis was chosen as time zero by TCGA, because this date was available for all patients but other dates, for example the date of sample procurement was not. In fact, date of sample procurement was not a required field. In addition, since dates are protected health information (PHI), some TSSs did not provide exact dates but only the month and year. For these cases TCGA BCR used the 15^{th} day of the month to represent the date. This best possible solution introduced a variation from 0 to 16 days for any such date which is in the same range of the days_to_sample_procurement for breast cancer. Thus, we concluded that date of diagnosis is the preferred choice for time zero in TCGA-CDR, when compared to date of sampling.

Definitions and derivation of clinical survival outcome endpoints

There are many definitions of clinical outcomes used in oncology research. After assessing all TCGA clinical datasets, four clinical survival outcome endpoints were chosen for this pan-cancer clinical endpoints analysis: Overall Survival (OS), Progression-Free Interval (PFI), Disease-Free Interval (DFI), and Disease-Specific Survival (DSS), defined as follows.

OS is the period from the date of diagnosis until the date of death from any cause. The censored time is from date of initial diagnosis until the date of last contact (largest number of days) from all the clinical data files (including both enrollment and follow-up forms). With minor exceptions as described in Handling of Special Cases and Problems in Clinical Data Files, derivation of OS outcomes was not problematic, as there exists minimal ambiguity in the databases about a patient's status at different time points: alive or dead.

PFI is the period from the date of diagnosis until the date of the first occurrence of a new tumor event (NTE), which includes progression of the disease, locoregional recurrence, distant metastasis, new primary tumor, or death with tumor. Patients who were alive without these event types, or died without tumor were censored (Hudis et al., 2007; Website). The event time is the shortest period from the date of initial diagnosis to the date of an event. The censored time is from the date of initial diagnosis to the date of last contact or the date of death without disease. PFI could be calculated for 31 of the 33 tumors, but not for acute myeloid leukemia (LAML) and lymphoid neoplasm diffuse large B cell lymphoma (DLBC). This endpoint is a commonly used surrogate endpoint for a future death outcome which otherwise takes a longer follow-up time to document and, unlike DFI, described below, it is associated with less ambiguity in that it is not necessary to first know whether a patient ever achieved disease-free status following their initial diagnosis and treatment. Nonetheless, several issues were encountered that needed resolution (c.f. Table S1 Tab TCGA-CDR_Notes). Among these, we needed to resolve the question of an NTE not clearly specified by tumor type, and to do this we adopted a more inclusive definition considering all NTEs in our calculation of PFI. However, in Table S1 Tab ExtraEndpoints, we provide two

more restrictive definitions of PFI (PFI.1 and PFI.2) so that users can choose the most appropriate PFI parameter for their research needs.

We would like to point out that the value of ''Progression of Disease'' was only reported in cancer types of OV, GBM, and UCS, for 34, 256, and 6 cases respectively. The definition of ''Progression of Disease'' was not as clear as the definitions for other PFI events, but they generally would be defined by clear radiographic evidence of new or progressive disease. OV and GBM were TCGA pilot study cancer types and that as the clinical data forms evolved and continued to be improved, the value of ''Progression of Disease'' was essentially eliminated from the data forms. At the launch of the full-scale TCGA study it had been replaced with more explicit conditions of new NTEs.

Importantly, in our definition of PFI, ''death with tumor'' was considered an event whether an NTE was reported or not. There were a total of 2478 patients who were ''Dead'' and ''With Tumor.'' For 1830 patients there was a defined NTE, but for 648 patients there was no recorded NTE. Since NTE was not a required field, these 648 patients may represent cases where an NTE was not recorded or that the disease rapidly progressed and the patient died without an NTE diagnosed. Counting the deaths without an NTE as PFI events would be expected to bias the outcome analysis due to overestimation of event times if the patient had developed an NTE prior to death but it was not reported to TCGA. To address whether these 648 patients represent under-reported NTEs and should be excluded from PFI studies we compared their overall survival time to patients who also died with tumor but who had an NTE reported. Because censored cases are excluded from this comparison and survival times tend to follow a log-normal distribution, we compared groups using the Wilcoxon rank-sum test (Table S2). Of the 31 cancer types that had available data for this analysis, 14 showed a shorter median event time for cases not having an NTE, of which 2 had significantly different survival distribution. The remaining 17 showed the opposite finding, of which 3 had significantly different survival distributions. Thus, in summary, only 5 cancer types (COAD, GBM, HNSC, KIRC, SKCM) had a significantly different survival-time distribution between the two groups of patients, with both over- and under-estimation observed. We therefore accept that there is not evidence for systematic bias by the assumption that in patients who were reported as dead with tumor be counted as PFI events and we retain these cases in analyses going forward.

DFI is another commonly used surrogate signifying future cancer mortality in many clinical studies. DFI defined here is the period from the date of diagnosis (due to the reason given below) until the date of the first new tumor progression event subsequent to the determination of a patient's disease-free status after their initial diagnosis and treatment. Such a new event can be either locoregional recurrence, distant metastasis, development of a new primary tumor in the same organ, or death from advancing of the same tumor (Hudis et al., 2007; Punt et al., 2007). Patients who developed a new primary tumor in another organ, or were alive without locoregional recurrence, without distant metastasis or development of another primary tumor in the same organ, or who were dead and tumor free were censored. The event time is the shortest time from initial diagnosis date to the date of an event. The censored time is from initial diagnosis date to the last contact date or the date of death.

This outcome endpoint was the most difficult to derive from available TCGA clinical data files in the absence of unspecified certainty about whether a patient was ever disease-free after their diagnosis. Consequently, 1095 stage IV TCGA patients were excluded from this endpoint analysis as recommended in other studies (Huo et al., 2017; Keenan et al., 2015) (Table S1 Tab TCGA-CDR), and given an NA (not applicable) for DFI. For other cases, the data field of ''Tumor Status'' in the initial enrollment data file could not be counted on, as it was unclear to us whether the clinical sites completing this field consistently followed a specific time point when the data form was completed, or if this indication was referring to the fact that the patient was once disease-free after the initial surgery or after the first course of treatment. We found cases supporting all possible scenarios. In certain cases, the evidence supported the situation that the patient had more than one-round of disease-free and recurrent cycles. After reviewing all the data elements we determined that the fields of ''treatment_outcome_first_course,'' ''residual_tumor,'' and ''margin_status'' could be used to arrive at this determination, and these fields were populated for 29 out of 33 cancer types (Table S4). Disease-free was defined as true, if the field ''treatment_outcome_first_course'' is ''Complete Remission/Response,'' the field ''residual_tumor'' is R0, or the field ''margin_status'' is negative. For SARC, the only disease that presented data of both ''margin_status'' and ''residual_tumor,'' where cases having both values were highly consistent, we choose ''residual_tumor'' to resolve conflicts as there were more cases having a value for ''residual_tumor'' than for ''margin_status.'' Also, in clinical practice the time to assess the ''residual_tumor'' is when first course of treatment was done. If the cases were never disease-free, they are given an NA (not applicable) for DFI. Cases of ''dead with tumor'' but without a new NTE were also excluded (given DFI of NA) as otherwise those cases would artificially prolong the time of recurrence and bias the results. LAML had only OS data, and SKCM, thymoma (THYM), and uveal melanoma (UVM) didn't have any of the information of ''treatment_outcome_first_course,'' ''residual_tumor,'' and ''margin_status,'' and their DFI were not available. Thus finally we were able to derive DFI for 5,521 cases (1,118 events and 4,403 censored) from 29 of the 33 tumor types (Table S4; Table S1 Tab TCGA-CDR).

We recognize that for DFI the time interval should start from the time when the patient was first determined to be disease-free, but such information was not available in the TCGA clinical dataset so we used the time of diagnosis as a surrogate. In statistical analysis using DFI, we restricted the patient set to those surviving at least 90 days to provide a proxy for the time required for a patient to complete treatment and be identified as disease free, to avoid immortal-time bias (Anderson et al., 1983; Giobbie-Hurder et al., 2013).

While OS is easy to define, it lacks specificity about cause of death and includes many non-cancer deaths. DSS, on the other hand, defined here as death from the diagnosed cancer type, has much greater relevance to cancer biology and therapeutic impact. A DSS event is death from the disease, and the event time is from the date of initial diagnosis until the date of death from the disease. The

censored time is from the date of initial diagnosis until the date of last contact or until the date of death from another cause. Within the TCGA dataset, however, derivation of DSS was complicated since only 6 of the 33 cancer types included a clinical form data field, "Cause of Death" (Table S3). Three of these cancer types, cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), pancreatic adenocarcinoma (PAAD), and UVM, had relatively high quality data for DSS analysis. Testicular germ cell tumors (TGCT) and prostate adenocarcinoma (PRAD) had few DSS and overall deaths. While, stomach adenocarcinoma (STAD) had over 50% of death events of unknown cause. For other cancer types, combining the fields of "tumor_status" with "vital status" allowed us to derive a surrogate for DSS, by approximating "Dead" and "With Tumor" as a DSS event, which in most cases was likely true. Note that the initial patient enrollment form data entry "Tumor Status" should not be considered as a standardized data field as it was poorly defined in early versions of the data form. The quality of the information in this field was improved by merging with the information from the follow-up data files, which yielded a new "tumor_status" field shown in Table S1 Tab TCGA-CDR, which we considered suitable for use in deriving this endpoint. But, we caution that this surrogate uni-directionally augments the number of events, as a patient with a tumor who died of other causes (e.g., a car accident or heart attack) would be counted as a DSS event. Thus our derived DSS only approximates the true DSS.

We would also like to point out that there is another often used endpoint, PFS. The definition of PFS is not standard and often relies on the clinical study at hand. In our definition of PFI, the events included deaths with tumor, but do not include deaths from other causes. We hold the opinion that for cancer studies, deaths with tumor are more relevant thus we choose to use our definition here which we term as PFI to tell its difference from commonly used term of PFS which, in some definitions, contains death from other causes. For convenience of users who prefer to use this definition of PFS, we included this endpoint in Table S1 Tab ExtraEndpoints.

Assessment of clinical endpoints of OS, PFI, DFI, DSS

Time-to-event studies must have sufficient follow-up to capture enough events and thereby ensure there is sufficient power to perform appropriate statistical tests (Clark et al., 2003). To test the sufficiency of the follow-up time for each endpoint, three tests and one supplemental check were applied in a two-step process. In Step 1, testing methods for sufficient follow-up developed by Maller and Zhou (Maller and Zhou, 1994) (Test 1), and by Shen (Shen, 2000) (Test 2) were used, employing a threshold of 0.05 for calculated statistic as significance. Test 1 is based on the concept that a plot of the K-M empirical distribution function tends to level off near its right extreme if the follow-up is sufficient. Thus, the step function of K-M should not jump at censored observations, and it should stay constant or flat on the interval from the largest failure/event time to the largest follow-up time. The magnitude of the interval is manifested by the information of the censored observations in the interval which is relevant to testing the hypothesis that the right extreme of failure/event distribution is less than or equal to the right extreme of the censoring distribution. If the observed value of the interval is "large," then the hypothesis is accepted so the study has a sufficient follow-up time. Under certain circumstances, however, this test may accept a hypothesis with a type I error larger than the nominal value of 0.05. Test 2 adopts the same concept as Test 1 but improved the control of the type I error. It uses the ratio of the largest follow-up time versus the largest failure/event time rather than the distance of the interval to perform the test. If the ratio is "large," then the hypothesis is accepted. Both tests, however, have the weakness that when applied to rapidly progressing diseases such as GBM, the test results are not always stable. To address this weakness, we proposed Test 3, requiring an event rate $\geq$ 30% for all the cases to complement Test 1 and Test 2.

In Step 2, we developed a Supplemental Check that was composed of three checkpoints: (1) a visual inspection of the cumulative event plot where the plot should reach a plateau (Figure S2); (2) median event time should be less than median censored time, to ensure relatively long follow-up time for events to occur but this condition is not necessary for diseases with an event rate > 50% such as GBM and ovarian cancer (OV); and (3) the number of events should be greater than 20 based on the "one in ten rule" in model building which means that one predictive variable can be studied for every 10 events (Harrell et al., 1996; Peduzzi et al., 1996). Thus, to approve an endpoint for use, the data should pass at least one of the 3 tests in Step 1, and pass all the checkpoints in the Step 2, the Supplemental Check. Data passing at least one of the three tests in Step 1 but failed to pass all three checkpoints in Step 2 yet had at least 10 events are recommended for use with caution. Otherwise the data are not recommended for use. The endpoints assessment method is illustrated in the flowchart Figure 1A.

The 4 survival end-points were tested for each of the 33 tumor types. For OS, among the 33 tumors, 7 tumor types passed both Test 1 and Test 2 in Step 1, which are adrenocortical carcinoma (ACC), BLCA, KIRP, rectum adenocarcinoma (READ), SARC, UCEC, and UVM, and further passed the Supplemental Check in Step 2 except for READ which may need a longer follow-up. 15 tumor types passed Test 1 (Maller and Zhou 1994), which are BRCA, CESC, COAD, ESCA, KICH, kidney renal clear cell carcinoma (KIRC), LGG, LUAD, OV, PCPG, PRAD, STAD, thyroid carcinoma (THCA), THYM, and uterine carcinosarcoma (UCS). Given that Test 1 is very conservative, we followed the Supplemental Check in Step 2 and confirmed that BRCA, LGG, PCPG, PRAD, and THCA did need a longer follow-up time to capture more OS events. GBM, HNSC, LUSC, MESO, PAAD, SKCM did not pass Tests 1 and 2 in Step 1, but passed Test 3, and further passed the Supplemental Check in Step 2, thus these 5 tumor types were approved as having a sufficient follow-up time for OS. Other diseases that did not pass the tests either needed a longer follow-up time or a larger sample size for more events. Thus OS of 21 tumor types was recommended for use without reservation, OS of 7 tumor types was recommended for use with caution, and for the remaining 5 tumor types OS was not an endpoint recommended for use.

For PFI, among the 33 tumors, LAML did not have the data. 19 tumor types passed Tests 1 and 2 in Step 1, which were BRCA, CESC, cholangiocarcinoma (CHOL), COAD, ESCA, KICH, KIRP, liver hepatocellular carcinoma (LIHC), LUAD, OV, PAAD, PRAD,

SARC, STAD, TGCT, THCA, THYM, UCEC, and UVM). They further passed the Supplemental Check in Step 2 except for BRCA and KICH; KICH had 17 events, and BRCA had a median time to event of 26 months and a median time to censor of 25 months. Since these two median times were so close plus all other conditions were met, we considered the follow-up time for PFI of BRCA as sufficient. 7 tumor types (DLBC, GBM, HNSC, LGG, LUSC, MESO, READ) passed Test 2, and 2 tumor types (ACC, and UCS) passed Test 1, and all of them passed the Supplemental Check in Step 2 except DLBC that had 12 events. 3 tumor types (BLCA, KIRC, and SKCM) did not pass Tests 1 or 2 but passed Test 3, and they further passed the Supplemental Check in Step 2. Thus, LAML did not have the PFI data; KIRC, DLBC and PCPG were recommended for use with caution, either because the sample size was too small or that a longer follow-up was needed to capture 20 events, and we also marked CHOL and UCS for use with caution because the sample sizes were small. The rest of the 27 tumor types were recommended for use without reservation.

The sufficiency of follow-up for DSS was similar to that of OS, and the sufficiency of follow-up for DFI was similar to that of PFI, except that the numbers of events were smaller.

Competing risk

The assumption of censoring in the above definitions of PFI, DFI, and DSS is that if the patient had not died they would have eventually experienced the event of interest. For instance, if a patient had not died of a heart attack he would have eventually died of his cancer. In this example death from a heart-attack is a competing risk for death from cancer. Or similarly, death from a heart attack is a competing risk for developing progression of cancer. Survival analysis models are available to account for these multiple event types and should be considered (Fine and Gray, 1999). The effect of a predictor on the outcome of interest may differ by outcome type. Modeling with competing risks may uncover these differences or reduce biases in the risk estimation from incorrect censoring assumptions. In the CDR a competing risks status variable is created for DSS, DFI, and PFI to assist with these analyses (Table S1 Tab ExtraEndpoints). In these status variables, "1" indicates that the event of interest occurred, "2" indicates that a competing event occurred, and "0" indicates that the patient did not experience an event during follow-up. For this paper, the influence of competing risks on the predictors under study was assessed and appeared to be minimal. For the high/low stage comparisons of Figures 2E-2G, considering this competing risk and found that the logHR was in the same direction, with the same significance, and similar magnitude. The difference in the two logHRs was very small and in both directions (Table S1 Tab Figure 2EFG_AdditionalInfo).

In deriving these endpoints, 11 data elements from the main or follow-up clinical data tables were used. We added another field to indicate whether the case was flagged for redaction based on the TCGA sample annotations.

Other clinical data fields processed

In addition to endpoint-related data fields, we also processed 10 commonly used data fields across cancer types wherever possible as it would be extremely difficult, inefficient, and beyond the scope of this pan-cancer effort, for us to process all clinical data fields given the different specific requirements for different cancer types. In these 10 common fields, quality assurance was performed by checking the data range and the logical relationships between data fields, and by comparison with previously derived data versions being used by other TCGA AWGs (Huo et al., 2017; Cancer Genome Atlas Network, 2012).

QUANTIFICATION AND STATISTICAL ANALYSIS

Cox proportional hazards (PH) regression model was used to calculate the Hazard Ratio (HR), the 95% confidence interval (95%CI), and p values, with the PH assumption assessed by a test of Schoenfeld residuals. The K-M method was used to create the survival plots and the log-rank test was used to compare the difference of survival curves. Wilcoxon rank-sum test was used for testing the difference between the distributions of un-censored survival time. Competing risks regression was used to estimate the competing risk of the endpoints per tumor type. For all tests, a two-tailed p value < 0.05 was considered statistically significant. In situations of multiple tests, the false discovery rate (FDR) was calculated using the Benjamini & Hochberg method. All analyses were performed using R 3.2.2.

DATA AND SOFTWARE AVAILABILITY

The TCGA clinical data were downloaded from the data portal of Genomic Data Commons (GDC, https://gdc-portal.nci.nih.gov/legacy-archive/). From the total of 225 files of TCGA clinical data, 130 initial enrollment and follow-up files were used, and a total of 11,160 patients across 33 tumor types were analyzed in this paper. An Integrated TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR) was created and available in Table S1.

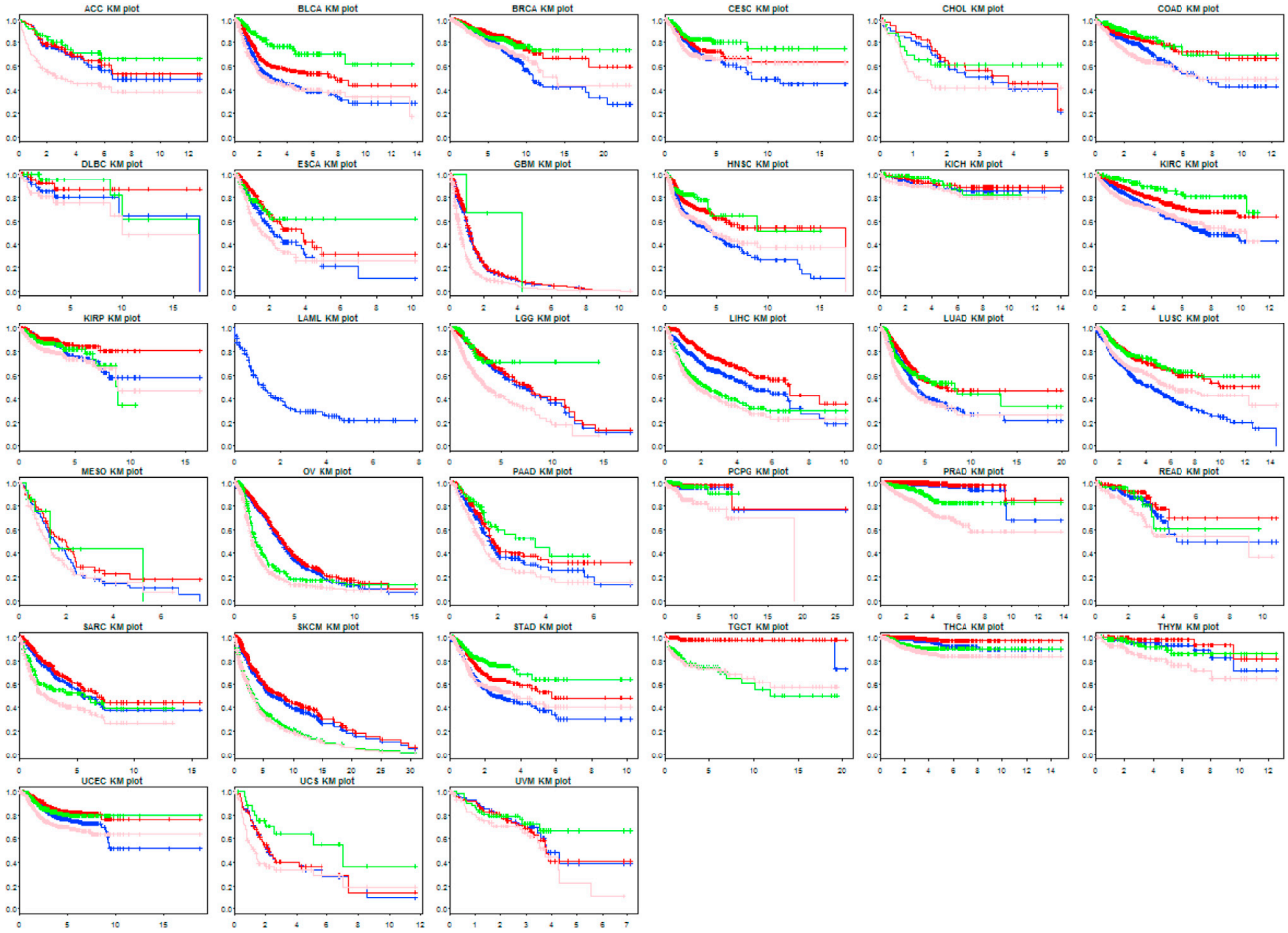


Figure S1. K-M Plots of OS, DSS, DFI, and PFI for 33 Tumor Types (OS, Blue; DSS, Red; DFI, Green; PFI, Pink), Related to Figure 1

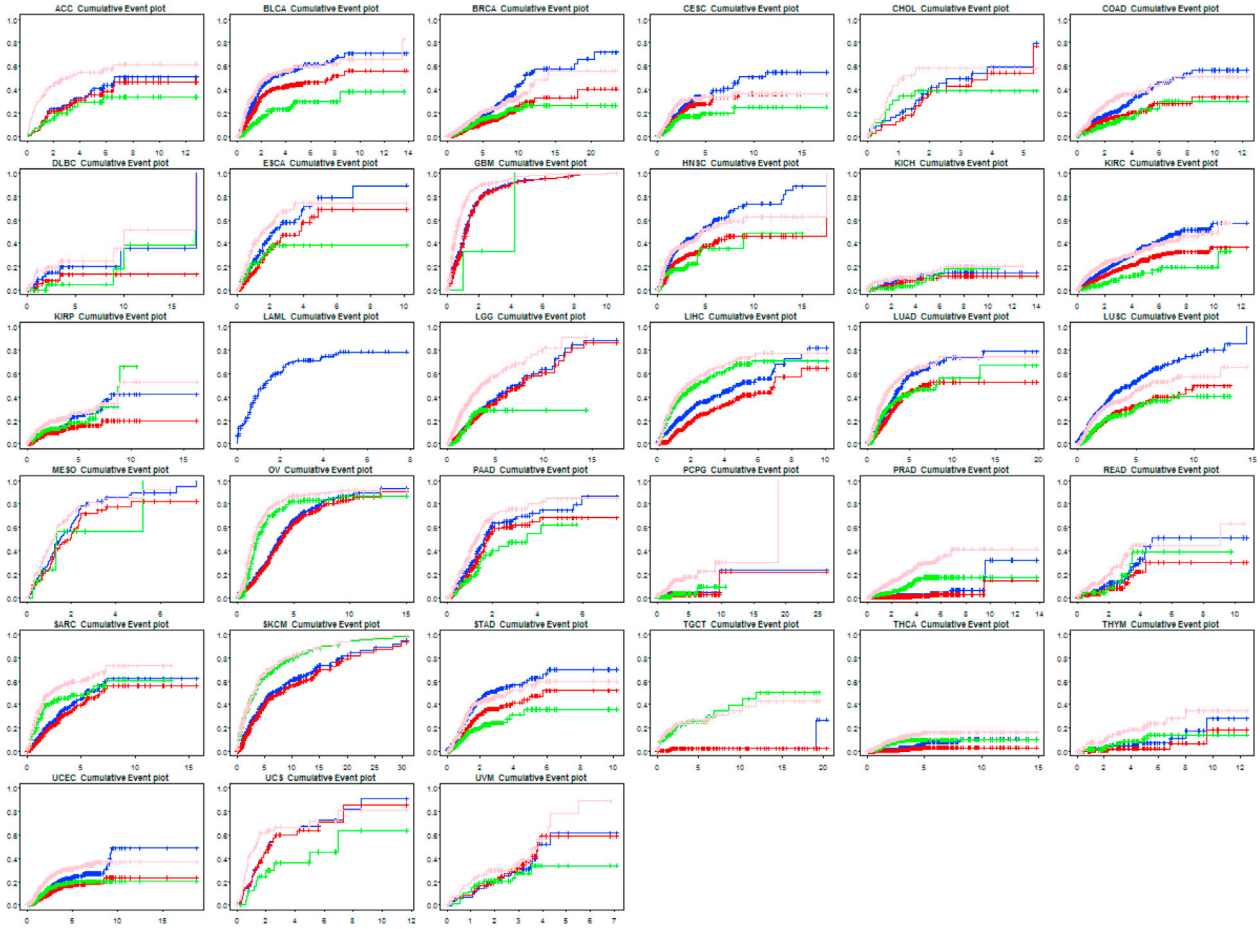


Figure S2. Cumulative Event Plot of OS, DSS, DFI, and PFI for 33 Tumor Types (OS, Blue; DSS, Red; DFI, Green; PFI, Pink), Related to Figure 1

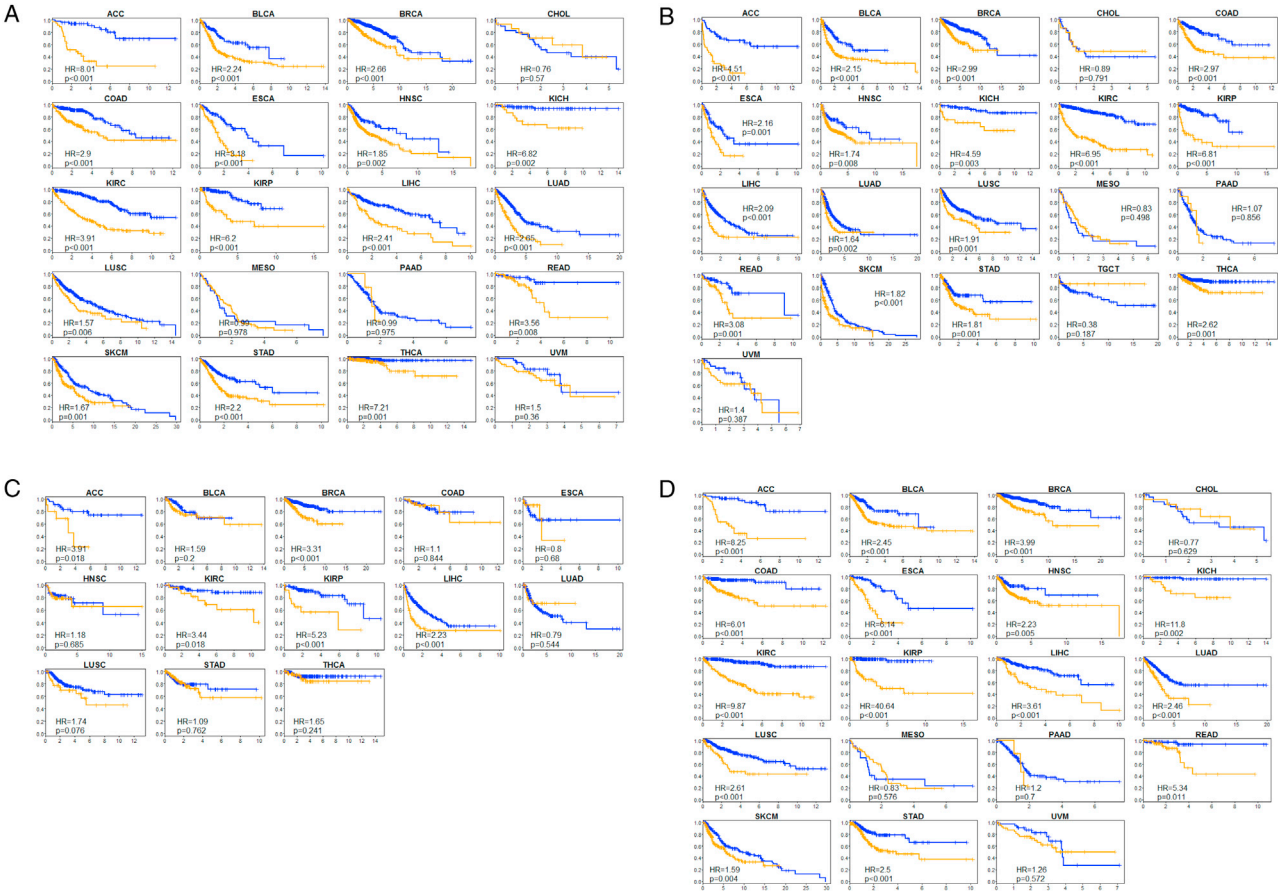


Figure S3. K-M Plots Comparing the Outcomes for Patients Diagnosed with Higher Stages (III and IV, Orange) versus Lower Stages (I and II, Blue), Related to Figure 2

(A–D) OS (A), PFI (B), DFI (C), and DSS (D). Only converge models were shown here.

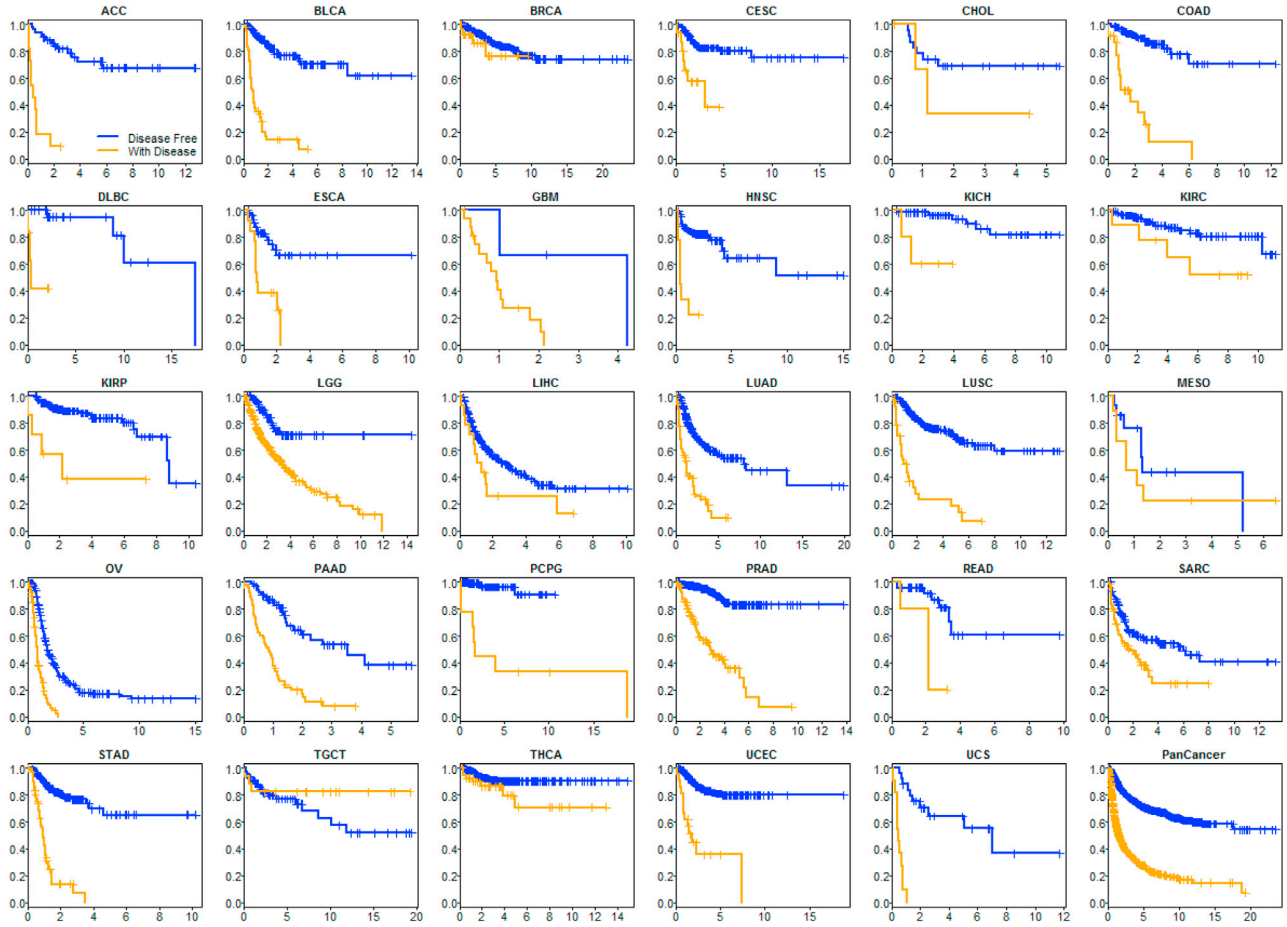


Figure S4. K-M Plots of new NTE Development from Patients Who Were Never Disease-Free (Orange) Compared with Those Who Were Once Disease-Free (Blue), Related to Table 4

Results of 29 cancer types are shown here as for the rest 4 of the 33 cancer types there was no new NTE information. The plot for all cancers combined is shown as the last figure.

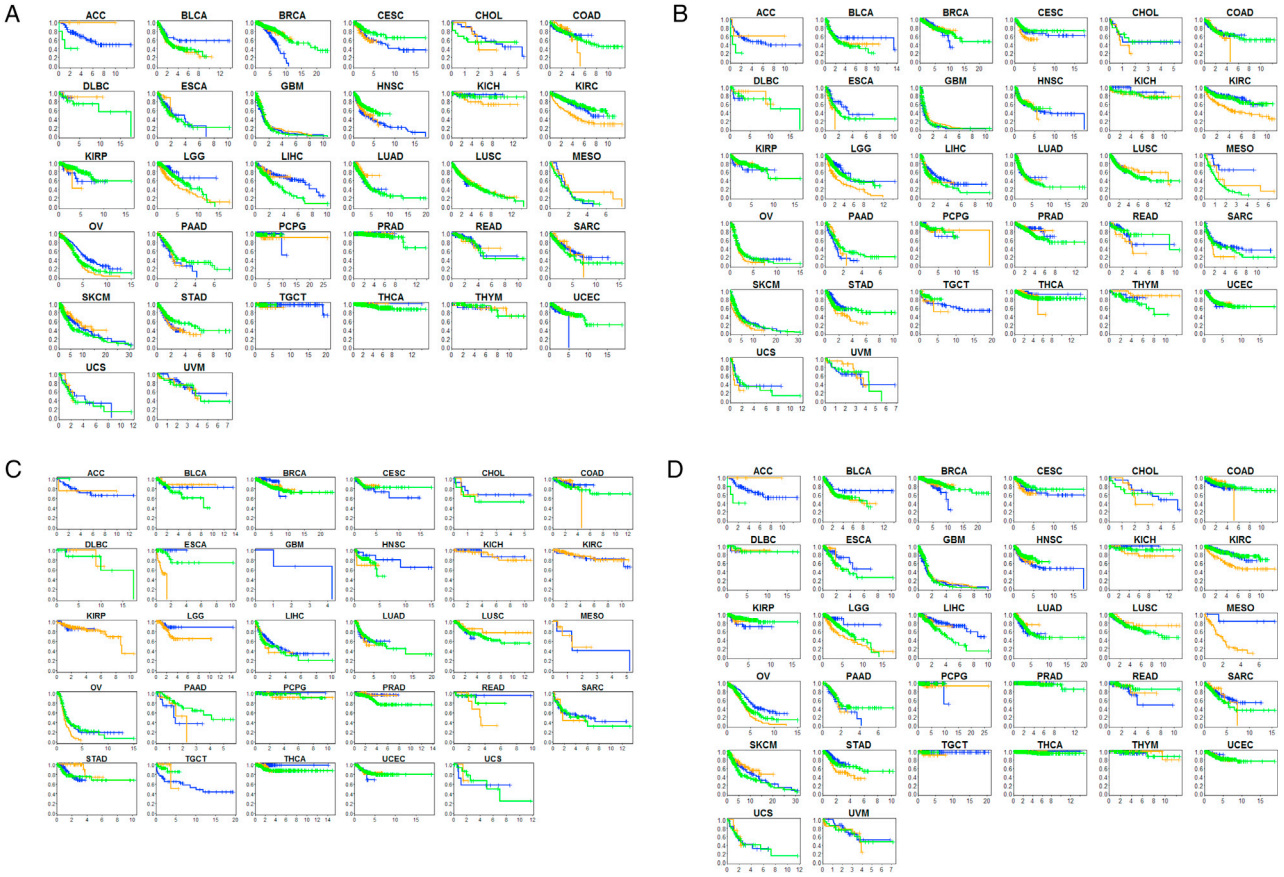


Figure S5. K-M Plots Comparing the Survival Data from the Top Two TSSs (#1 in Blue and #2 in Orange) with Those from All Other Sites (Green) for Each Cancer Type, Related to Table 5

(A–D) Plots were made using endpoints of OS (A), PFI (B), DFI (C), and DSS (D). Statistical analysis results for diseases with both top sites supplying at least 50 cases are shown in Table 5.