DO CANCER CLINICAL TRIAL POPULATIONS TRULY REPRESENT CANCER PATIENTS? A COMPARISON OF OPEN CLINICAL TRIALS TO THE CANCER GENOME ATLAS

NOPHAR GEIFMAN

Institute for Computational Health Sciences, University of California, San Francisco
Mission Hall, 550 16th Street
San Francisco, CA 94158-2549, USA
Email: Nophar.Geifman@ucsf.edu

ATUL J. BUTTE

Institute for Computational Health Sciences, University of California, San Francisco
Mission Hall, 550 16th Street
San Francisco, CA 94158-2549, USA
Email: Atul.Butte@ucsf.edu

Open clinical trial data offer many opportunities for the scientific community to independently verify published results, evaluate new hypotheses and conduct meta-analyses. These data provide a springboard for scientific advances in precision medicine but the question arises as to how representative clinical trials data are of cancer patients overall. Here we present the integrative analysis of data from several cancer clinical trials and compare these to patient-level data from The Cancer Genome Atlas (TCGA). Comparison of cancer type-specific survival rates reveals that these are overall lower in trial subjects. This effect, at least to some extent, can be explained by the more advanced stages of cancer of trial subjects. This analysis also reveals that for stage IV cancer, colorectal cancer patients have a better chance of survival than breast cancer patients. On the other hand, for all other stages, breast cancer patients have better survival than colorectal cancer patients. Comparison of survival in different stages of disease between the two datasets reveals that subjects with stage IV cancer from the trials dataset have a lower chance of survival than matching stage IV subjects from TCGA. One likely explanation for this observation is that stage IV trial subjects have lower survival rates since their cancer is less likely to respond to treatment. To conclude, we present here a newly available clinical trials dataset which allowed for the integration of patient-level data from many cancer clinical trials. Our comprehensive analysis reveals that cancer-related clinical trials are not representative of general cancer patient populations, mostly due to their focus on the more advanced stages of the disease. These and other limitations of clinical trials data should, perhaps, be taken into consideration in medical research and in the field of precision medicine.

1. Introduction

Approximately 30,000 clinical trials are conducted each year across the globe and various market and regulatory forces are driving initiatives to publicly share patient-level data from these trials [1-3]. With the advancement of science and betterment of the human condition in mind, there are several purported benefits for the sharing of clinical trial data [4-6]. Sharing these data offers the opportunities for the scientific community to independently verify published results. Lack of availability of original research data is a known, significant, barrier against reproducibility. Availability of the data may provide opportunities to evaluate new hypotheses that were not originally formulated in the studies, either by extending the analysis of data from a clinical trial or by combining data from different trials.

The availability of clinical data from different trials makes such data an attractive source for systemic research and meta-analysis [7]. Examining disease-related patterns by meta-analysis can help gain better understanding of disease-related characteristics and lead to new discoveries and insights. Combining data from multiple clinical studies and evaluating the same disease with various outcome measures could help leverage an improvement in efficacy by suggesting possible combination of treatments. Additionally, these data may be used to identify and define subgroups of subjects who respond better to a specific treatment. The plethora of raw, individual-level clinical data should provide a real springboard for scientific advances in precision medicine and development of new techniques in clinical informatics [8-10].

Due to the complexity of the different types of cancer, and the difficulties in selecting the right therapeutic approaches, increasing efforts are being dedicated towards improving cancer care via precision medicine [10, 11]. While most of the focus in this field is on the genetics and molecular characteristics of the cancer and increasing a drug efficacy based on the properties of a given tumor, other forms of clinical data can be useful in advancing precision medicine in cancer.

Recently, several pharmaceutical clinical trial data sharing platforms such as the Project Data Sphere [1, 2] and the *clinicalstudydatarequest.com* site (developed by GlaxoSmithKline) have emerged, making raw data from clinical trials available for research. Another, well-established source for cancer-related clinical data is The Cancer Genome Atlas (TCGA), which aims to comprehensively characterize and analyze many cancer types and makes its data freely available for research [12]. While the TCGA's focus lies in the genetics of cancer, establishing a large database of cancer genome sequences and aberrations, it also holds clinical data such as patient survival, treatment and demographics.

While pharma-released data is increasingly becoming available for research, to our knowledge there are very few works that utilized these data sources. Here we present the integration of patient-level data from many cancer clinical trials, present the potential of these data and systematically evaluate whether, in the field of cancer, trial data can usefully represent patient populations.

2. Methods

2.1. *Clinical trial data from the Project Data Sphere portal*

Clinical trial data was obtained from the Project Data Sphere portal (projectdatasphere.org) which stores, shares and allows analysis of patient level phase III cancer trial data [2]. The database currently holds data from the comparator arms of 48 cancer clinical trials. Subjects assigned to the comparator arms of cancer clinical trials usually receive standard of care treatment. Following registration, submission and approval of a research proposal, the data is made available for either download or analysis on an online SAS platform (Figure 1).

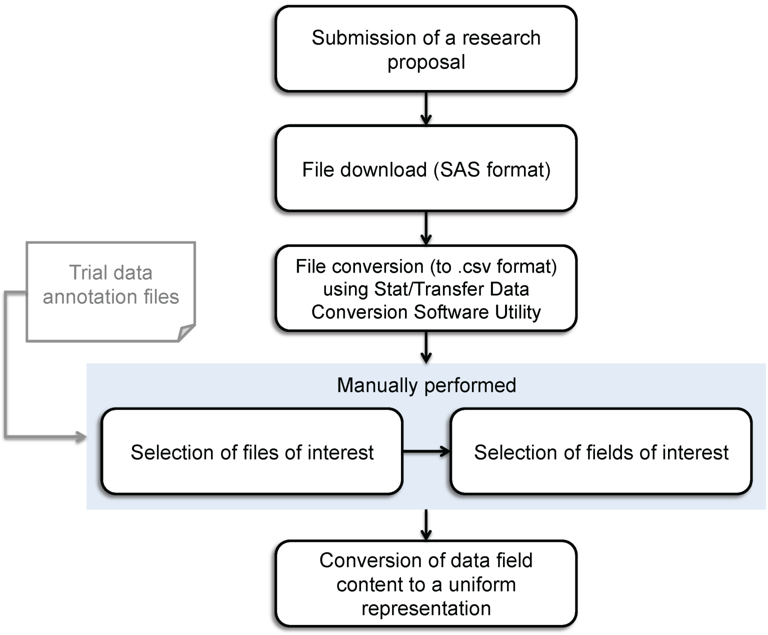
Submission of a research proposal
File download (SAS format)
File conversion (to .csv format) using Stat/Transfer Data Conversion Software Utility
Trial data annotation files
Manually performed
Selection of files of interest
Selection of fields of interest
Conversion of data field content to a uniform representation

Fig. 1. Clinical trial data processing workflow.

Survival, demographic and treatment data for 10 cancer clinical trials (the first made available on the site which matched cancer types in the TCGA and for which annotation files were available) were downloaded from the Project Data Sphere database and used for the analyses presented here (Table 1).

Table 1. Clinical trial data downloaded from the Project Data Sphere portal.

	Cancer type	Study sponsor	No. of subjects*	ClinicalTrial.gov ID
1	Breast	Pfizer	166	NCT00373113
2	Colorectal	Astra Zeneca	690	NCT00384176
3	Prostate	Astra Zeneca	635	NCT00626548
4	Prostate	Astra Zeneca	248	NCT00672282
5	Prostate	Astra Zeneca	550	NCT00673205
6	Prostate	Centocor	45	NCT00385827

7	Prostate	Celgene	226	NCT00988208
8	Prostate	Pfizer	278	NCT00676650
9	Prostate	Sanofi	371	NCT00417079
10	Pancreas	Sanofi	156	NCT00417209

The majority of subjects have stage IV cancer. Approximately half (n=1439) of the prostate cancer subjects from the Astra Zeneca trials do not have a specific stage assigned. All breast cancer subjects are female; all prostate cancer subjects are male. In colorectal cancer, there are 290 females and 400 males. In the pancreatic cancer dataset, there are 64 females and 92 males. * The number of subjects for which survival data was available.

2.2. *The Cancer Genome Atlas dataset*

Data from the TCGA was selected to represent the general cancer subject population, representing a wide range of cancer stages, patient ages, ethnic groups and treatments. For each of the four cancers evaluated in this study (breast, colorectal, prostate and pancreatic), survival, treatment and demographic data were obtained from the TCGA database, April 27, 2015. These cancer types were selected due to their matching to cancer types available in our clinical trial dataset (Table 2). Only subjects for which survival data was available were included. Subcategories of cancer stages were converted to a more inclusive stage, for example: stages IIa, IIb, IIc were converted to stage II.

Table 2. The Cancer Genome Atlas dataset.

Cancer type	No. of subjects*	Gender (%)	Stage (%)
Breast	1018	Male - 12 (1.2) Female - 1006 (98.8)	I - 179 (17.6) II - 574 (56.4) III - 239 (23.5) IV - 21 (2.1) NA - 5 (0.5)
Colorectal	355	Male - 190 (53.5) Female - 165 (46.5)	I - 55 (16.74) II - 136 (38.8) III - 107 (28.4) IV - 48 (14.1) NA - 9 (2)
Prostate	492	Male - 492 (100) Female - NA	II - 337 (68.5) III - 48 (9.8) IV - 4 (0.8) NA - 103 (20.9)
Pancreas	171	Male - 94 (55) Female - 77 (45)	I - 18 (10.5) II - 142 (83) III - 5 (2.9) IV - 4 (2.3) NA - 2 (1.2)

* The number of subjects for which survival data was available.

2.2.1. *Assigning cancer stage for the prostate cancer subset*

For the prostate cancer data subset, the cancer stage was not available for any of the subjects. We therefore assigned each subject with a cancer stage based on the tumor and metastasis status for which data was available. If the assigned metastasis status is M1, stage IV is assigned to that subject (regardless of tumor status). For metastasis status M0 and tumor status T1b, T1c, T1 and T2, stage II was assigned. For metastasis status M0 and tumor status T3, stage III was assigned and for metastasis status M0 and tumor status T4, stage IV was assigned. There were no other combinations of tumor and metastasis status in the data.

2.3. *Survival analyses*

Kaplan-Meier analyses were conducted using the ‘Survival’ package (version 2.38-3) in R [13]. For calculation of the significance for the difference between sample sets, the log-rank test was used. For each analysis (by cancer type, cancer stage, gender etc.) subjects for which data was missing (from either dataset) were excluded.

3. Results

3.1. *Clinical trial subjects show lower survival in comparison to TCGA subjects*

We first compared the survival of subjects between the four different cancer types (breast, colorectal, prostate and pancreatic) in each of the datasets. The survival of subjects from the clinical trial dataset are overall significantly ($p<0.0001$) lower in comparison to matching cancer types from the TCGA (Figure 2). This observation can logically be explained by the generally more progressive cancer stage of subjects included in the trial dataset, in comparison to the TCGA. One interesting observation however, is the differences in the order (from best to worst survival) of cancers in the survival plots, illustrated in Figure 2A and 2B. In trial subjects, breast cancer shows lower survival than colorectal while in TCGA subjects, breast cancer has a higher survival than colorectal cancer. In both datasets, the differences in survival between breast and colorectal cancer subjects are significant ($p<0.005$). Comparison of the demographics of trial and TCGA subjects reveals that age distribution is similar in both datasets other than for the “over 75 years” age group (Figure 3A). For breast, colorectal and pancreatic cancers, the TCGA has more subjects over the age of 75 than the clinical trials; indicating that, clinical trials tend to select fewer older subjects. Gender distribution is similar in both datasets (Figure 3B) and the majority of subjects in both datasets are Caucasian (Figure 3C).

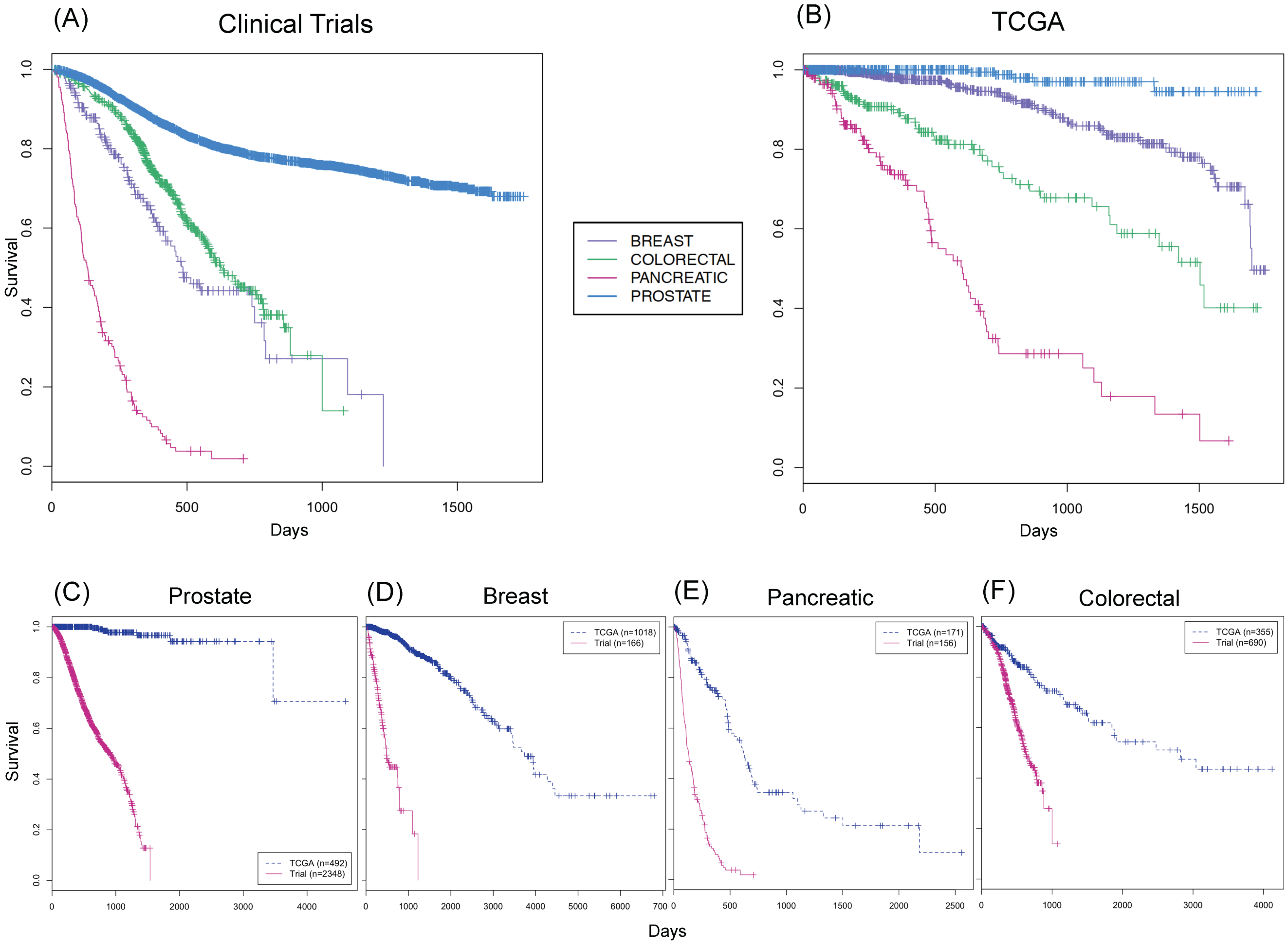

Fig. 2. Cancer patient survival in clinical trials and TCGA.

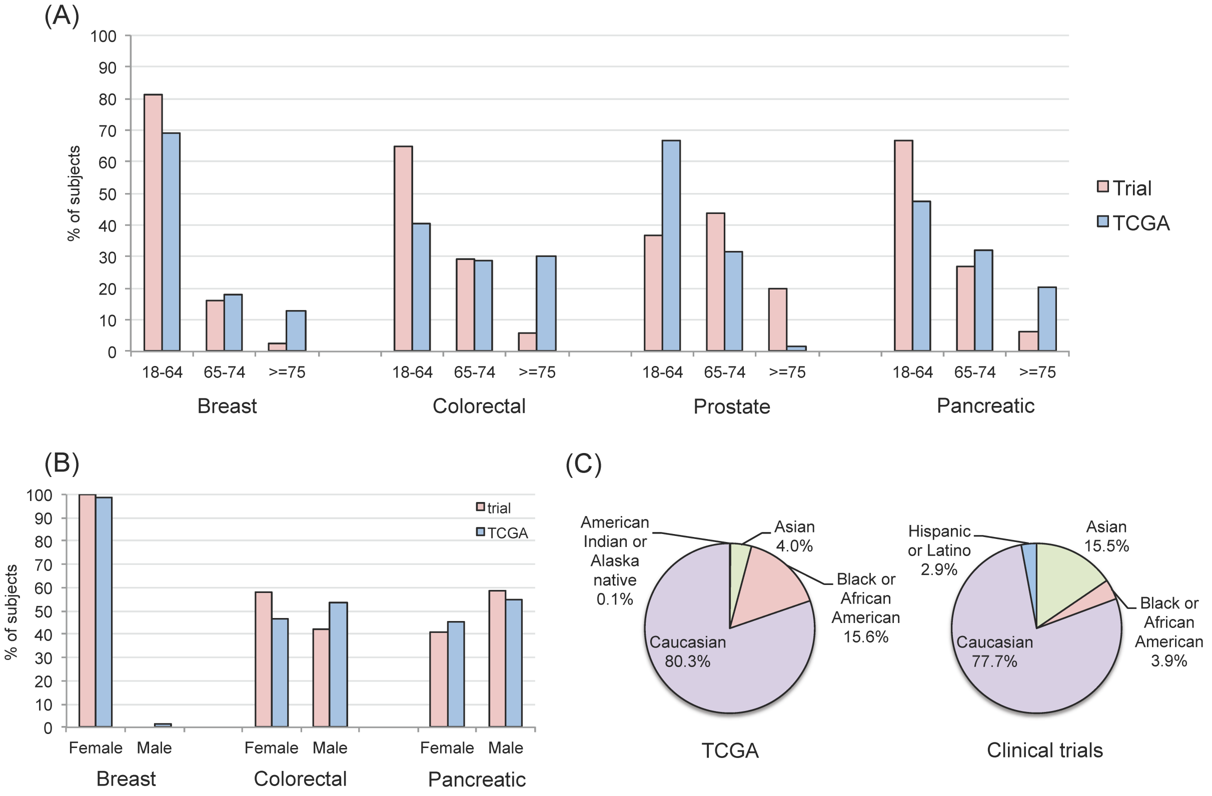

Fig. 3. Demographic variables in clinical trials and TCGA. (A) Age distribution in each cancer subset. (B) Gender distribution in each cancer subset (prostate cancer was excluded since subjects are always male). (C) Ethnic distribution in each of the datasets.

3.2. *Gender differences in clinical trial data but not TCGA*

We next compared the survival of subjects from trials to those from the TCGA on the basis of gender. When comparing subjects from all cancer types (breast, colorectal, prostate and pancreatic), significant differences in survival were found between males and females in the trial dataset (p=3.37e-12) but not in the TCGA (Figure 4). To evaluate whether it is the gender-specific cancer types (such as breast or prostate) which are driving this difference, we next compared the survival between genders in each dataset but this time limiting the analysis to only the pancreatic and colorectal cancer types which have a similar distribution of males and females in both datasets. In this instance, no significant gender differences in survival were found in either of the datasets.

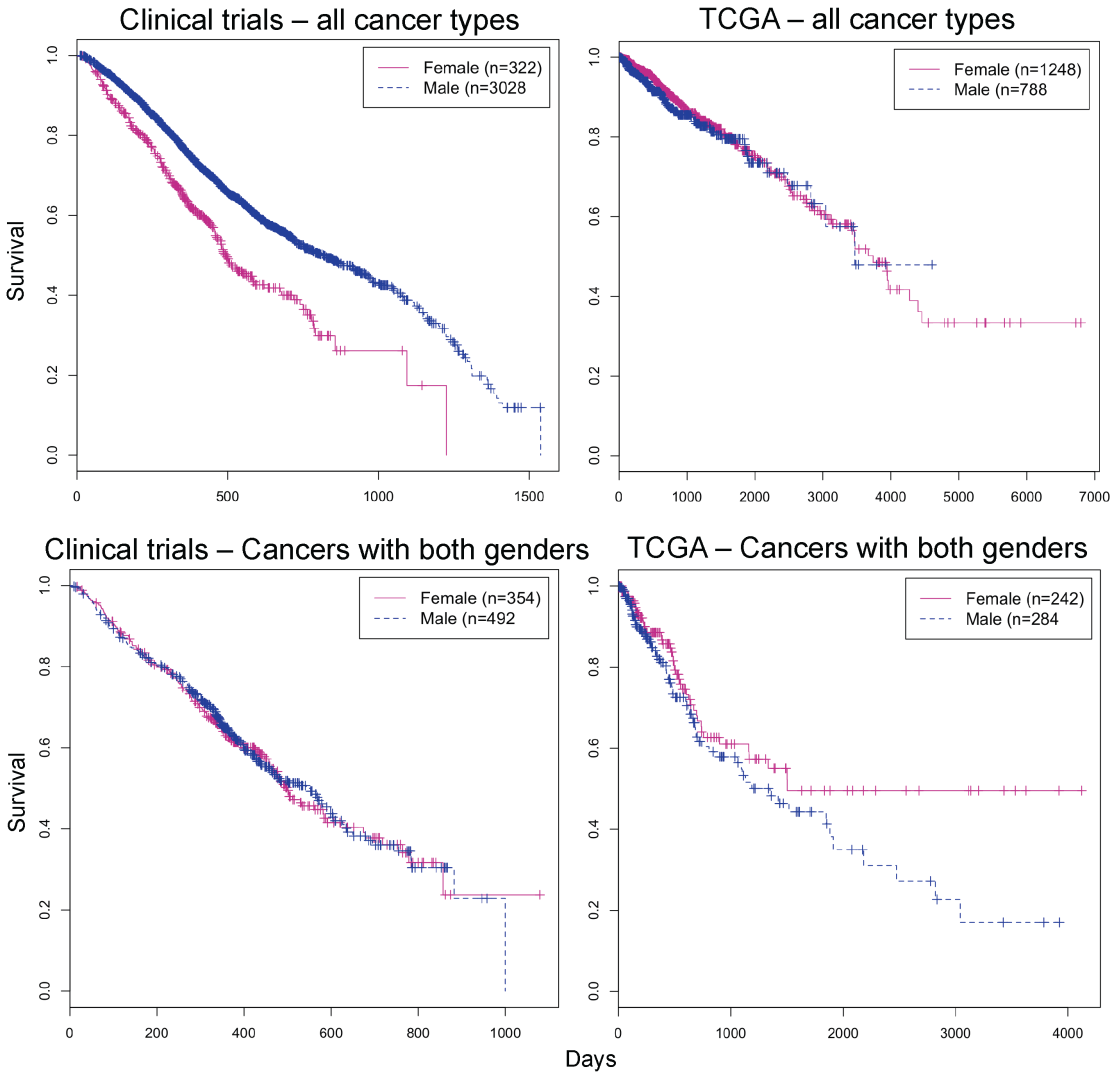

Fig. 4. Gender differences in survival in clinical trials and TCGA.

3.3. *The effect of the cancer stage*

In order to evaluate the differences in survival based on patients' cancer stages in the TCGA and trial datasets, for each cancer type all available stages from the TCGA were compared to stage IV subjects from the trial dataset (Figure 5). The trial data was limited to only stage IV (or advanced) subjects since that was the only available stage for most of the cancer types (breast, colorectal and pancreatic) in this dataset. In all four cancer types, stage IV subjects from clinical trials showed lower survival rates than subjects with other stages of cancer from the TCGA. In addition, in all four cancer types, subjects with stage IV cancer from the trials dataset showed lower survival than stage IV subjects from TCGA; significantly in breast cancer ($p<0.005$), though for colorectal cancer this was not significant ($p=0.764$) and for pancreatic and prostate cancer ($p<0.05$) the TCGA sample size of stage IV subjects was very small, see Table 2).

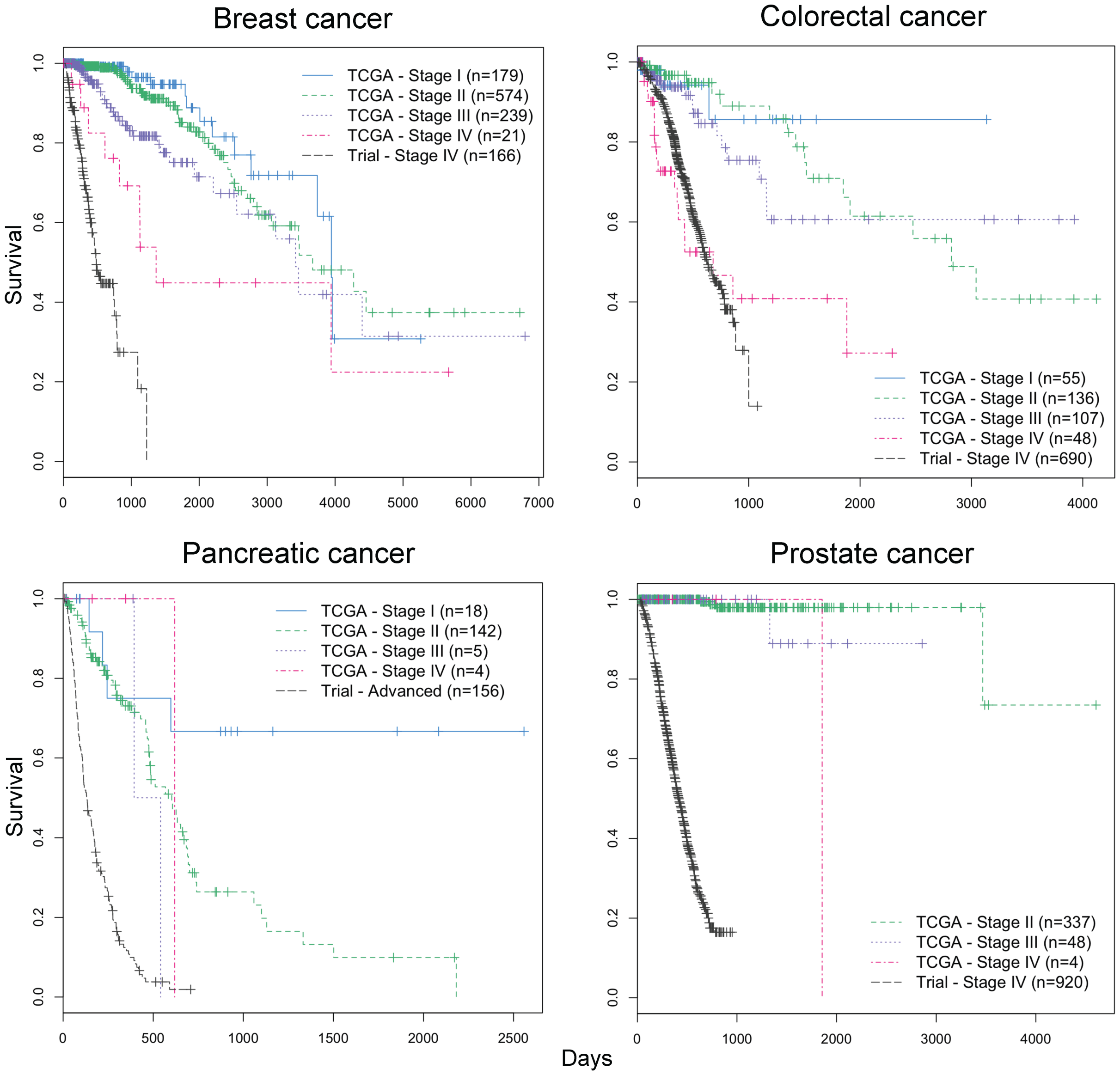

Fig. 5. Survival of different cancer stages in TCGA and clinical trials.

4. Discussion

Recently, clinical trial data sharing platforms such as the Project Data Sphere, the *clinicalstudydatarequest.com* site (developed by GlaxoSmithKline) and ImmPort [14]; have been making raw data from clinical trials available for investigation. These emerging, rich datasets offer many opportunities for research and the advancement of precision medicine. To evaluate whether cancer-related pharma-released trial data is representative and comparable to cancer patient populations, we present here the comparison of cancer-related patient-level data from a newly open clinical trial dataset to The Cancer Genome Atlas.

Our first analysis, comparing survival of different cancer types between the two datasets, revealed that clinical trial breast cancer subjects show lower survival than clinical trial colorectal cancer subjects; while in the TCGA dataset, breast cancer has a higher survival than colorectal cancer. Overall, lower survival of trial subjects in comparison to TCGA subjects can be explained by the generally more advanced stage of cancer in trial subjects. However, differences in cancer stages cannot explain the differences between the datasets in survival of breast and colorectal cancer. In both these cancers in the trial dataset, all subjects are stage IV and in both those cancers in the TCGA datasets, subjects have a mix of stages with no significant differences between the distribution of cancer stages in breast and colorectal subjects.

Based on this observation, it can be deduced that for stage IV cancer, colorectal cancer patients have a better chance of survival than breast cancer patients. On the other hand, for all other stages, breast cancer patients have better survival than colorectal cancer patients. Generally speaking, trial subjects have a worse prognosis; this is unlikely because they are in a trial, but rather because of how they are selected for trials. It should be pointed out that this observation is made based on a single breast cancer trial and a single colorectal cancer trial. Therefore, it is possible that other factors, such as practices of the company conducting the trial, differences in tumor subtypes or the population’s characteristics may contribute to the differences in survival between these cancer types. Further analysis, which would include data from additional breast and colorectal cancer trials, could further elucidate this observation.

Comparison of survival by gender revealed that survival differences between genders exist in trial data, but not in the TCGA dataset. This is most likely driven by the stage of the cancer rather than the gender of the subjects. More specifically, by the lower survival of breast cancer subjects in the trial dataset in comparison to that of prostate cancer subjects. Since in the trial data, all breast cancer subjects are stage IV while the prostate cancer subjects are of a mix of stages (I, II, II and IV) it could be the cancer stage which is the cause for the difference. This gender difference is not seen in the TCGA dataset, probably because both the breast and prostate cancer subjects have a similar distribution of a mix of stages (with very few stage IV subjects).

Subjects with stage IV breast, pancreatic and prostate cancer from the trials dataset showed lower survival than matching stage IV subjects from TCGA. One possible explanation for this is that subjects included in clinical trials tend to have undergone several lines of therapy and only because those have failed, have these subjects enrolled in a trial. On the other hand, many of the subjects in the TCGA dataset are facing their first line of treatment. Therefore, the colorectal cancer subset is more comparable, since inclusion criteria for the colorectal cancer trial was that subjects have not undergone previous treatment for colorectal cancer and that the study’s treatment is the first line of treatment. For this cancer type, no significant differences were found between the survival of trial and TCGA subjects; further indicating that stage IV trial subjects have lower survival rates since their cancer is less likely to respond to treatment.

One of the major shortcomings our analysis is that while we were able to compare the two datasets based on cancer type, cancer stage or gender, one factor that was not taken into consideration is the treatment given to the subjects. Matching data from TCGA to the that of a respective clinical trial dataset while taking into consideration all possible variables including treatment (type and course) is virtually impossible, likely to leave a very small number of subjects if any at all. For example, in the TCGA there are only 4 stage IV colorectal cancer patients who were treated with FOLFOX and would be comparable to the colorectal trial subjects. On the other hand, there is indication that there are few treatment-associated differences in the survival or long-term outcomes of early or localized prostate cancer subjects [15, 16]. Therefore, it could be argued that the survival differences shown here between TCGA and trial subjects in prostate cancer may be caused, at least in part, by something other than differences in treatment. As more data becomes available, integrative analyses taking treatment arms into account can be carried out and meta-analysis can be conducted across many clinical trials.

The clinical trials dataset we describe here holds a lot of potential for research. However, while getting hold of the data was a fast and relatively straightforward process, other than when using the somewhat restrictive online tools, the data were far from readily usable. One of the major issues we encountered was that for the available studies, data schemas are not uniform and table and variable names were lacking annotation which had to be manually extracted from long annotation files. Since considerable manual effort needed to be invested in the processing of these data, only ten trials (of the 48 available) were downloaded and used as proof of concept for the work presented here. Future work will include adding data from all trials available on the Project Data Sphere site. In addition, as the trend of opening clinical trials data to the research community grows, further clinical trial datasets can be included. While the results of the work described here indicate that clinical trial data does not accurately represent subject populations (for the cancer types evaluated here), this dataset can still be very useful. The availability of clinical data from different trials makes such data an attractive source for systemic research and meta-analysis. For example, one possible use, given a large enough number of studies pertaining to the same disease, would be to conduct a meta-analysis and compare the outcomes of different treatments. Examining disease-related patterns by meta-analysis can help gain better insight into disease-related characteristics and assist in finding new discoveries and insights. Moreover, by combining data from multiple clinical studies we can leverage the improvement in efficacy by possible combination of treatments.

5. Conclusions

We present here the analyses of data from a newly available pharma-released clinical trial dataset and its comparison to data from The Cancer Genome Atlas. Our comprehensive analysis reveals that clinical trials, in the field of cancer, are not representative of cancer patient populations, probably due largely to their focus on advanced stages. However, while recognizing that these data have their limitations, they hold great potential for advancing medical research in the field of precision medicine. Future research design should include consideration of the representativeness of the cancer patient population.

6. Acknowledgments

The authors would like to thank Dr. Marina Sirota and Dr. Dvir Aran for helpful discussion and useful suggestions.

7. Disclaimer

This publication is based on research using information obtained from www.projectdatasphere.org, which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC nor the owner(s) of any information from the web site have contributed to, approved or are in any way responsible for the contents of this publication.

References

1. A. K. Green, et al., The Project Data Sphere Initiative: Accelerating Cancer Research by Sharing Data. *Oncologist* (2015).
2. K. Hede, Project data sphere to make cancer clinical trial data publicly available. *J Natl Cancer Inst* **105** 16 (2013).
3. P. Nisen and F. Rockhold, Access to patient-level data from GlaxoSmithKline clinical trials. *N Engl J Med* **369** 5 (2013).
4. H. M. Krumholz and E. D. Peterson, Open access to clinical trials data. *JAMA* **312** 10 (2014).
5. B. Lo, Sharing clinical trial data: maximizing benefits, minimizing risk. *JAMA* **313** 8 (2015).
6. M. Rosenblatt, S. H. Jain and M. Cahill, Sharing of clinical trial data: benefits, risks, and uniform principles. *Ann Intern Med* **162** 4 (2015).
7. J. S. Ross, R. Lehman and C. P. Gross, The importance of clinical trial data sharing: toward more open science. *Circ Cardiovasc Qual Outcomes* **5** 2 (2012).
8. in *Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease*. 2011: Washington (DC).
9. N. V. Chawla and D. A. Davis, Bringing big data to personalized healthcare: a patient-centered framework. *J Gen Intern Med* **28 Suppl 3** (2013).
10. A. R. Shaikh, et al., Collaborative biomedicine in the age of big data: the case of cancer. *J Med Internet Res* **16** 4 (2014).
11. L. A. Garraway, J. Verweij and K. V. Ballman, Precision oncology: an overview. *J Clin Oncol* **31** 15 (2013).
12. N. Cancer Genome Atlas Research, et al., The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45** 10 (2013).
13. T. T. A Package for Survival Analysis in S. version 2.38. (2015); Available from: http://CRAN.R-project.org/package=survival.
14. S. Bhattacharya, et al., ImmPort: disseminating data to the public for the future of immunology. *Immunol Res* **58** 2-3 (2014).
15. M. R. Cooperberg, J. M. Broering and P. R. Carroll, Time trends and local variation in primary treatment of localized prostate cancer. *J Clin Oncol* **28** 7 (2010).
16. L. Holmberg, et al., A randomized trial comparing radical prostatectomy with watchful waiting in early prostate cancer. *N Engl J Med* **347** 11 (2002).