Impact of Clinical Data Veracity on Cancer Genomic Research

Sunali Mehta, PhD 1,2,† Deborah Wright, MBChB, PhD, FRACS 1,2,† Michael A Black, PhD 1,4,† Ahmad Anjomshoaa, PhD 3, Fran Munro, MSc 6, John McCall, MBChB, FRACS, MD 1,3, Cristin Print, MBChB, PhD 1,3

1Maurice Wilkins Centre, University of Auckland, Auckland, New Zealand; 2Pathology Department, University of Otago, Dunedin, New Zealand; 3Department of Surgical Sciences, University of Otago, Dunedin, New Zealand; 4Department of Biochemistry, University of Otago, Dunedin, New Zealand; 5Department of Surgery, University of Auckland, Auckland, New Zealand; 6Medical Genetics Division, Faculty of Medicine, Kerman University of Medical Sciences, Kerman, Iran; and 7Department of Molecular Medicine and Pathology, University of Auckland, Auckland, New Zealand

†Authors contributed equally to this work.

Correspondence to: Cristin Print, MBChB, PhD, Department of Molecular Medicine and Pathology, University of Auckland, Private Bag 92019, Auckland, New Zealand (e-mail: c.print@auckland.ac.nz).

Abstract

Genomic analysis of tumors is transforming our understanding of cancer. However, although a great deal of attention is paid to the accuracy of the cancer genomic data itself, less attention has been paid to the accuracy of the associated clinical information that renders the genomic data useful for research. In this brief communication, we suggest that omissions and errors in clinical annotations have a major impact on the interpretation of cancer genomic data. We describe our discovery of annotation omissions and errors when reviewing an already carefully annotated colorectal cancer gene expression dataset from our laboratory. The potential importance of clinical annotation omissions and errors was then explored using simulation analyses with an independent genomic dataset. We suggest that the completeness and veracity of clinical annotations accompanying cancer genomic data require renewed focus by the oncology research community, when planning new collections and when interpreting existing cancer genomic data.

Translational research combining clinical data with genomic data [eg, (1-3)] has dramatically increased our understanding of cancer and lays the foundation for precision oncology (4,5). Clinically annotated cancer genomic datasets from individual laboratories (4), biobanks (6), or large international initiatives (1,7) can now be accessed without bioinformatics expertise using web interfaces such as cBioPortal (8), Xena (9), or StratomeX (10). However, as clinical and genomic data were integrated by large-scale initiatives, inaccuracies were recognized (11), and unexpected issues such as hidden duplicates emerged (12). In clinical databases and registers in general, clinical data errors are surprisingly pervasive (13-16), including missing clinical annotations that reduce analysis power (17). Despite the considerations described above, relatively little has been done to evaluate the effects of clinical data omissions and errors on the interpretation of cancer genomic datasets.

When a senior clinician reviewed the clinical data accompanying a colorectal cancer gene expression dataset from our laboratory, we discovered a substantial number of omissions and errors, despite the high degree of care initially taken when manually extracting data from the contemporaneous clinical record. Data review using clinical notes, direct communication with medical staff, and publicly available records such as death certificates identified 1 or more missing data points in 142 of 205 (69%; Figure 1, A) tumor records and errors in 32 of 205 (16%; Figure 1, B) tumor records. Additional clinical data were identified that clarified metastatic status for 103 of the 205 (50%) tumors and, on review, the American Joint Committee on Cancer stage (7th edition) was revised for 12 tumors, and the designation of recurrence status was corrected for 21 tumors, which impacted on disease-free survival (DFS) for 16 of 205 (8%) tumors. The changes made are summarized diagrammatically in Figure 1, C, and the overall effect on DFS is shown in Figure 1, D. As an example, the impact of these corrections to clinical data on conclusions drawn from downstream genomic analysis was assessed using Affymetrix gene expression data from these 205 tumors. Cox proportional hazards analysis of associations between transcriptome-wide gene expression and DFS identified 79 RNAs with statistically significant associations with DFS following revision.

To estimate the impact of omissions and errors in clinical data fields associated with genomic datasets in general, we...
Figure 1. Review and correction of the clinical annotations of a colorectal cancer gene expression dataset. Clinicopathological data pertaining to the 205 colorectal tumor samples that were collected prospectively in New Zealand between 1996 and 2013 is compared before and after clinical review and correction. 

A) Graphic illustrating the overall relationships between different types of clinical data omissions or errors, and the corrections they required. White bars indicate a data point missing in the original dataset. Pale shades indicate a data point that was correct in the original data set and was unmodified by revision. Dark shades indicate an inaccurate data point in the original dataset, which was corrected in the revised dataset. Grey bars in the DFS column indicate a stage IV tumor not included in calculation of DFS. The black bar crossing all columns indicates a neuroendocrine tumor excluded from subsequent revision or analysis. Beneath this graphic is the number and % of missing and inaccurate data of each type in the original clinical annotation.

B) Kaplan-Meier DFS analysis is shown for the original clinical data (blue) and the revised data (red) for the same patient cohort. The x-axis shows time since diagnosis in years; the y-axis shows proportion of patients remaining disease free. A log-rank test indicated that the difference between these survival curves is statistically significant ($P = 0.005$). Samples and data have been obtained with informed consent and ethical approval from the New Zealand multicenter ethics committee. DFS = disease-free survival; OS = overall survival; M = metastasis; N = lymph nodes; T = tumor; TTR = time to relapse; stage = American Joint Committee on Cancer stage (7th ed).
Figure 2. 1000-fold simulation analysis of the effect of missing and incorrect clinical annotations on gene expression-DFS associations for colorectal tumors. The Gene Expression Omnibus GSE17536 gene expression data set was used (18). Cox proportional hazards survival models were used with a cut-off for statistical significance of a $P$ value less than .05 without (A, C, E) and with (B, D, F) Benjamini-Hochberg multiple testing correction. Within each panel, boxplots indicate the distribution of the number of RNAs significantly associated with DFS (left set of boxes), RNAs that gained de novo statistically significant associations with DFS after the simulated changes (middle set), and RNAs that lost their statistically significant associations (right set). Various combinations of percentage of patients with data changes and...
underwent a simulation analysis using a publicly available colorectal cancer gene expression dataset (GSE17536, n = 177) (18). The effect of clinical data omissions and errors on significant associations between gene expression and DFS was determined using Cox proportional hazards analysis with a statistical significance cut-off of a P value less than .05, with and without multiple testing correction. Fixed proportions of clinical annotations were omitted or changed 1000 times in randomly chosen patients. The results confirm that omissions or changes to relatively small subsets of high-impact clinical data have a substantial impact on RNAs associated with DFS (Figure 5). For example, omission of clinical annotations for 20% of patients caused almost all previously statistically significant (P ≤ .05) RNAs associated with DFS to be lost and reduced sensitivity for detecting survival associations by one-third on average, when compared with the original data set (Figure 2, A and B). Changes to disease status (Figure 2, C and D) and time to death (Figure 2, E and F) also had a substantial impact on the sets of genes with statistically significant association with DFS and on sensitivity. For example, random switching of survival status for only 10% of patients caused more than 50% of the RNAs associated with DFS to be lost (Figure 2, C and D).

Missing or erroneous clinical information accompanying genomic datasets can originate either in the original clinical record or during the extraction, transcription, or interpretation of the data (15). Our study, data was collected from 1996 to 2013, with a period (2000-2010) of substantially reduced clinician oversight. Secondly, several recurrent tumors were found on expert review to be second primary cancers. This is particularly important when investigating associations between gene expression and behavior of the incident tumor. Thirdly, substantial changes in clinical practice occurred during this study, including the introduction of routine pre-operative computed tomography staging, which may have led to stage migration during the study (19). Similar iterative improvements to imaging and pathology technologies (20) are likely to affect many genomic studies. Changes to the type and accuracy of clinical data recorded electronically may also have led to ascertainment biases (16,21), as may the more readily available survival data for patients who have died compared with patients who are still alive but temporarily lost to follow-up (22).

Lessons we have learned from this exercise include clear documentation of any systematic changes in clinical practice over time, providing embedded clinician oversight at the point of data entry, following standardized training for all data collection staff, creating logically arranged synoptic data collection forms, censoring of the survival status of all patients at the same point in time, and undertaking periodic expert data reviews.

We suggest that the effects of omissions and errors may similarly impact more nuanced associations between clinical and genomic information, such as associations between mutational profiles and the efficacy of targeted therapies. Our findings may not be surprising to statisticians, however, many clinicians and cancer biologists who regularly use genomic data are blissfully unaware of these effects. We suggest to our colleagues that 1) it is worth considerable investment to ensure that clinical annotations of genomic data are as complete and accurate as possible, especially when genomic resources will be used in population-scale analyses (23), developing clinical tests, or in linked clinical-genomic workflows (24) and 2) ideally, standardized sets of clinical variables should be included in all clinical metadata. Our conclusions also have implications for interpretation of large international cancer genomic projects such as The Cancer Genome Atlas or International Cancer Genome Consortium, as well as for smaller scale genomic analyses in individual laboratories. In addition, the revision of clinical annotations may change the conclusions drawn from already published cancer genomic studies.

Funding

This work was supported by the New Zealand Foundation for Research Science and Technology (CP and SM), The Health Research Council of New Zealand (DW), the Maurice Wilkins Centre (CP), and The Colorectal Surgical Society of Australia and New Zealand (DW, AM, CP).

Notes

Role of the funder: The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the article; or the decision to submit the article for publication.

Disclosures: The authors have no conflicts of interest to disclose.

Author contributions: Writing-original draft and Writing-review & editing SM, DW, MB, AM, AA, FM, AR, JM, CP; Conceptualization, Formal Analysis and Funding acquisition SM, DW, CP; Methodology and Software SM, MB; Supervision CP.

Acknowledgements: We wish to thank and acknowledge work of the University of Otago Microarray Facility and the Dunedin Colorectal Cohort (https://blogs.otago.ac.nz/dncrc).

Disclaimers: The authors have no disclaimers to declare.

the degree of these changes are shown. Simulations shown in A and B estimate the effect of data omission, C and D changes to alive or dead patient status, and E and F time to relapse. Lower and upper bounds of the boxes indicate the 25th and 75th percentiles, respectively. Horizontal lines in boxes indicate the 50th percentile. Lower and upper bounds of the whiskers indicate 5th and 95th percentile, respectively. For example in A, the leftmost [yellow] box plot represents number of RNAs significantly associated with DFS (P < .05) in the original dataset. The orange, red, and pink boxes represent total gain, or loss of RNAs significantly associated (P < .05) with DFS after 1000 simulated omissions of 5%, 10%, or 20% of data, respectively. Q summarizes the effects of simulated loss or gain of clinical data on the ability to detect statistically significant prognostic associations. Distributions of sensitivity (− recall; TP/[TP + FN]), specificity (TN/[TN + FP]), positive predictive value (PPV) − precision; TP/[TP + FP]), and negative predictive value (NPV; TN/[TN + FN]) are tabulated, where TP is true positive, TN is true negative, FP is false positive, and FN is false negative, with positive and negative defined as Cox proportional hazards association between RNA transcript abundance and DFS of a P value of no more than .05 or a P value of more than .05, respectively. The original GSE17536 gene expression data set is considered the gold standard for these simulations, in which sensitivity, specificity, PPV, and NPV would be considered to have a value of 1. Omission or change to components of clinical data is seen to affect sensitivity much more than specificity is impacted, because although similar numbers of statistical associations with DFS are gained as are lost in the simulations (shown in A-F), in the original data set, only a small proportion of genes have a statistically significant Cox proportional hazards association (P < .05) with DFS. These calculations used the confusion matrix function of the caret package (25) in the R statistical software framework. DFS = disease-free survival.
Data Availability

The study follows the FAIR principles, however the data set used in this analysis will be made available through the corresponding author, after consideration by a data access committee of its potential use. The data are not made freely available through international genomic data sharing initiatives due to ethical requirements of the studies in which this data was generated, which respect cultural considerations about genomic data of New Zealand’s indigenous Maori people.

References

1. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-337.  
2. Kazaninaeva M, Mehta S, Eiholzer RA, et al. The Δ133p53Δ isoform promotes an immunosuppressive environment leading to aggressive prostate cancer. Cell Death Dis. 2019;10(9):631.  
3. Lawrence B, Blenkiron C, Parker K, et al. Recurrent loss of heterozygosity correlates with clinical outcome in pancreatic neuroendocrine cancer. NPJ Genom Med. 2018;3:18.  
4. Lasham A, Knowlton N, Mehta SY, et al. Breast cancer patient prognosis is determined by the interplay between TP53 mutation and alternative transcript expression: insights from TP53 long amplicon digital PCR assays. Cancers (Basel). 2021;13(7):1531.  
5. Muthukaruppan A, Lasham A, Wood KJ, et al. Multimodal assessment of estrogen receptor mRNA profiles to quantify Estrogen pathway activity in breast tumors. Clin Breast Cancer. 2016;17(2):139-153. doi:10.1016/j.clbc.2016.09.001.  
6. Roelands J, Decock J, Boughorbel S, et al. A collection of annotated and harmonized human breast cancer transcriptome datasets, including immunologic classification. F1000Res. 2017;6:296.  
7. International Network of Cancer Genome projects. Nature. 2010;464(7291):993-998. doi:10.1038/nature08987.  
8. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cbioPortal. Sci Signal. 2013;6(269):pl1.  
9. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675-678.  
10. Streit M, Lex A, Gratzl S, et al. Guided visual exploration of genomic stratifications in cancer. Nat Methods. 2014;11(9):884-885.  
11. Yuan Y, Van Allen EM, Omberg L, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol. 2014;32(7):644-652.  
12. Waldron L, Rister M, Ramos M, et al. The doppelganger effect: hidden duplicates in databases of transcriptome profiles. J Natl Cancer Inst. 2016;108(11):djw146.  
13. Warsi A, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. Eur J Surg Oncol. 2002;28(8):850-856.  
14. Arts DG, De Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc. 2002;9(6):600-611.  
15. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. AMIA Annu Symp Proc. 2008;2008:242-246.  
16. Hong MK, Yao HH, Pedersen JS, et al. Error rates in a clinical data repository: lessons from the transition to electronic data transfer—a descriptive study. BMJ Open. 2013;3(5):e002406.  
17. Lung P, Zhong D, Pang X, et al. Maximizing the reusability of gene expression data by predicting missing metadata. PLoS Comput Biol. 2020;16(11):e1007450.  
18. Smith JJ, Deane NG, Wu F, et al. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology. 2010;138(3):958-968.  
19. Feinstein AR. On classifying cancers while treating patients. Arch Intern Med. 1985;145(10):1780-1791.  
20. Harris G, O’Toole S, George P, et al. Massive parallel sequencing of solid tumours—challenges and opportunities for pathologists. Histopathology. 2017;70(1):123-133.  
21. Renshaw AA, Gould EW. Impact of immediate access to the electronic medical record on anatomic pathology performance. Am J Clin Pathol. 2013;140(1):109-111.  
22. Bull K, Spiegelhalter DJ. Tutorial in biostatistics survival analysis in observational studies. Statist Med. 1997;16(9):1041-1074.  
23. Singal G, Miller PG, Agarwala V, et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non-small cell lung cancer using a clinicogenomic database. JAMA. 2019;321(14):1391-1399.  
24. Gruendner J, Wolf N, Togel L, et al. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR) system design and implementation. J Med Internet Res. 2020;22(10):e19879.  
25. Kuhn M. Building predictive models in R using the caret package. J Stat Soft. 2008;28(5):1-26.