Title

Interrater Reliability in Toxicity Identification: Limitations of Current Standards

Permalink

https://escholarship.org/uc/item/444177pp

Journal

International Journal of Radiation Oncology • Biology • Physics, 107(5)

ISSN

0360-3016

Authors

Fairchild, Andrew T
Tanksley, Jarred P
Tenenbaum, Jessica D
et al.

Publication Date

2020-08-01

DOI

10.1016/j.ijrobp.2020.04.040

Peer reviewed

Title: Inter-rater reliability in toxicity identification: Limitations of current standards

Running Head: Toxicity inter-rater reliability

Authors:
Andrew T. Fairchild, MD, MAT[1]
Jarred P. Tanksley, MD, PhD[1]
Jessica D. Tenenbaum, PhD[2]
Manisha Palta, MD[1]
Julian C. Hong, MD, MS[1,3,4]

1. Department of Radiation Oncology, Duke University, Durham, NC
2. Department of Biostatistics and Bioinformatics, Duke University, Durham, NC
3. Department of Radiation Oncology, University of California, San Francisco, San Francisco, CA
4. Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA

***Corresponding Author:**
Julian C. Hong, MD, MS
Department of Radiation Oncology and Bakar Computational Health Sciences Institute
University of California, San Francisco, San Francisco, CA

1825 Fourth Street, Suite L1101
San Francisco, CA 94158

E-mail: julian.hong@ucsf.edu

Conflict of Interest Notification:
JDT, MP, and JCH are coinventors on a pending patent, "Systems and methods for predicting acute care visits during outpatient cancer therapy," broadly but not directly related to this manuscript.

Key Words: Radiation therapy; chemoradiation; toxicity; standardization; CTCAE

Funding: None

Abstract

Purpose: The NCI Common Terminology Criteria for Adverse Events (CTCAE) v5.0 is the standard for oncology toxicity encoding and grading despite limited validation. We assessed inter-rater reliability (IRR) in multi-reviewer toxicity identification.

Methods and Materials: Two reviewers independently reviewed 100 randomly selected notes for weekly on-treatment visits during radiotherapy from the electronic health record (EHR). Discrepancies were adjudicated by a third reviewer for consensus. Term harmonization was performed to account for overlapping symptoms in CTCAE. IRR was assessed based on unweighted and weighted Cohen's kappa coefficients.

Results: Between reviewers, unweighted kappa was 0.68 (95% CI 0.65-0.71) and weighted kappa 0.59 (0.22-1.00). IRR was consistent between noted present or absent symptoms with kappa of 0.6 (0.66-0.71) and 0.6 (0.65-0.69), respectively.

Conclusions: Significant discordance suggests toxicity identification, particularly retrospectively, is a complex and error prone task. Strategies to minimize IRR, including training and simplification of the CTCAE criteria, should be considered in trial design and future terminologies.

Introduction

The NCI Common Terminology Criteria for Adverse Events (CTCAE) v5.0 is the current standard for oncology toxicity encoding and grading, though reliability studies and validation have been limited.[1] Inter-rater reliability (IRR) impacts both prospective and retrospective studies, given reliance on multiple individuals to assess toxicities. Prior data suggest variability between clinicians and research assistants in grading toxicities[2,3] and underreporting in clinical trials versus retrospective chart review.[4] Thus, retrospective chart review has been considered the gold standard and review by clinical research assistants is frequently employed to capture acute events. The increasing complexity of CTCAE across versions has also made acute event grading more challenging.[3] Variability of toxicity identification, necessary for accurate grading, has not been quantified; this may further reveal shortcomings of both toxicity identification and, more broadly, retrospective studies to guide future improvement.

Methods and Materials

This study was approved by the Duke University Medical Center Institutional Review Board (Pro00082776). Two senior radiation oncology residents independently reviewed 100 randomly selected notes for weekly on-treatment visits (OTVs) from the electronic health record (EHR) from 2005-2016. Patients undergoing radiotherapy have weekly visits to assess and manage potential acute toxicities of treatment. These notes can be institution-specific but typically include a brief subjective section describing patient symptoms, an objective section detailing a limited and focused physical exam, and a brief assessment and plan. At our institution, OTV notes are documented in the broader medical center EHR and free text comprises the majority of notes, although vital signs and physical exam headers are prepopulated and physical exam findings can be selected from predefined options. Style and content varied across physicians and disease site. At other institutions, OTVs may be documented using more structured formats with limited free text. OTV notes were selected given their overall simplicity in comparison to consultation or follow-up notes which tend to have significant text generated from the EHR. CTCAE v5.0 symptoms were identified as explicitly present, absent, or not mentioned. Reviewers were instructed to identify all mentioned symptoms and were blinded to each other's identifications. An attending radiation oncologist acted as a tie-breaker and evaluated disagreements to facilitate an overall consensus. We created a thesaurus to harmonize overlapping CTCAE terms (Supplementary Material). IRR was assessed based on unweighted and weighted Cohen's kappa coefficients between reviewers and versus the consensus.[5,6] The Kappa coefficient is a commonly used measure of IRR which accounts for agreement due to chance, and ranges from -1 to +1, where 0 is agreement expected from random chance and 1 is perfect agreement.[7] The weighted Cohen's kappa coefficient was calculated with greater weight given to symptoms where one reviewer designated a symptom as explicitly present, while the other identified it as explicitly absent. We additionally evaluated IRR separately for explicitly present and absent symptoms.

Results

One hundred notes written by 15 physicians and representing many disease sites revealed disagreements in symptom identification in 93 notes, with median 4 per note (range 1-12) (**Table 1). Figure 1** illustrates relationships between reviewers and the consensus with regard to present and absent symptoms. Radiation dermatitis and pain were among the most common documented symptoms. Certain terms were more thoroughly captured by individual reviewers, including non-infectious cystitis, folliculitis, and soft tissue fibrosis.

Between the blinded raters, inter-rater unweighted kappa was 0.59 (95% CI 0.56-0.62) and weighted kappa was 0.5 (-1.00-1.00). With harmonization, this improved, with unweighted kappa 0.68 (95% CI 0.65-0.71) and weighted kappa 0.59 (0.22-1.00). Consensus agreement with each reviewer was asymmetric, with unweighted kappa of 0.95 (0.93-0.96) and 0.75 (0.72-0.77), and weighted kappa of 0.93 (0.93-0.93) and 0.67 (CI 0.67-0.67). IRR was consistent between noted present or absent symptoms with kappa of 0.6 (0.66-0.71) and 0.6 (0.65-0.69), respectively.

Discussion

CTCAE remains the cornerstone of standardized oncology toxicity grading, with retrospective chart review playing an important role in acute toxicity identification. Nevertheless, significant discordance was present even with two practicing resident physicians (familiar with CTCAE and responsible for routine documentation) reviewing brief clinical notes with limited automatically populated content such as past medical history or social history from the EHR. This suggests that consistent identification, upstream of toxicity grading, is a challenging.[2,3] Moreover, variability may even be underestimated in this study as only 77 of 808 possible harmonized terms were present by consensus. This study demonstrates the limitations of retrospective toxicity review, despite its use as a gold standard in prior work and auditing due to underreporting in prospective studies.[4] Our findings also highlight general shortcomings of retrospective studies. As manual review can be laborious, many studies rely on chart abstraction from single reviewers and some have multiple reviewers abstracting data from different patients. Our findings emphasize the caution that is required in interpreting this abstracted data.

Multiple plausible factors contribute to discordance. In particular, the CTCAE has increased in complexity,[3] including 837 terms across 26 categories in v5.0, increased from 49 and 9, respectively, in v1.0. Perhaps not surprisingly, selecting the appropriate term can be problematic; rectal bleeding from prostate cancer radiotherapy may be coded in accurate terms such as “proctitis” or “rectal hemorrhage.” We developed and have made available a harmonization thesaurus to account for common symptoms encountered in this study with overlapping terms such as hematochezia and pain. The improvement in kappa with harmonization reinforces the potential redundancy of CTCAE terms and demonstrates, to a degree, how this impacts inter-rater variability. These results support a prior study which demonstrated that the increase in terms from CTCAE v1.0 to v4.0 resulted in a decreasing ability for clinical research assistants to conclusively grade acute events.[3]

This complexity may also limit the number of symptoms a human reviewer can actively recall during identification, reflected by the discrepancy in the identification of specific symptoms, such as non-infectious cystitis, folliculitis, and soft tissue fibrosis. Some inter-rater variability in our study may also stem from elements of ambiguity in clinical documentation, including those due to uncertain terms or chronological changes. CTCAE does not account for symptom changes over time;[5] thus a statement such as “sore throat earlier in the week that resolved” could be coded as both present and absent. Other terms such as “minimal” can also contribute to human variability.

Contributions of retrospective review to IRR are also important to consider in this study, given variability in documentation style and author identification. Additionally, our practice at our institution has been to rely heavily on unstructured free text OTV notes. Greater structure for reporting toxicities such as drop-down menus may improve future extraction, though the high complexity of CTCAE may make comprehensive collection difficult. However, while some retrospective series are based prospectively structured collected outcomes, this has been documented to result in greater underreporting than those based on retrospective review.[4] While our study focused on identifying toxicities described in OTV notes, larger studies often abstract data from even more complex consult and follow-up notes, which likely increase the cognitive burden on the reviewer and may result in even greater variability.

Finally, variations in the “completeness” of reviewer assessment, particularly with negated symptoms, impact the overall IRR. Explicitly positive symptoms or unambiguous categorical items, such as laboratory data, are abstracted with high fidelity.[1,4] More variability, though, occurs when inference is required, particularly in identifying explicitly negative symptoms; “OP/OC clear” during head and neck treatment plausibly relates to the absence of thrush and/or mucositis; “dysphagia is improved” is ambiguous regarding resolution versus a continued, albeit less severe, presence. Despite these challenges of CTCAE, though, there did not appear to be greater variability with explicitly present or absent symptom identification.

Our study has limitations, particularly small sample size and number of reviewers given the intensiveness of manual review of the expansive CTCAE terminology. While this does limit our ability to quantify the degree of IRR, it does reflect common research practice in both retrospective and prospective settings, where the burden of manual review frequently limits studies to one reviewer. Metrics for interrater reliability each have limitations, and the Kappa metric assumption of independence may underestimate agreement.[7] However, Kappa has the advantage of accounting for potential concordance due to chance, particularly important given that, as in many studies, much of the CTCAE terminology was not used. Prior experience in EHR data extraction in general nor CTCAE toxicity identification was not standardized amongst the reviewers, though neither reviewer had additional experience abstracting toxicities beyond clinical training of a senior resident. Reviewers were asked to identify all potential toxicities and did not receive dedicated CTCAE harmonization or delineated rules for ambiguous terms. Restricting the CTCAE or standardization efforts may improve IRR.[8] However, this reflects common practice and emphasizes that training may be an appropriate strategy in studies, particularly in multi-institutional and cooperative group settings.

Finally, the growing experience around patient-reported outcomes[9] and interest in EHR extraction with natural language processing[10] may offer complementary sources of data to augment the accurate identification of acute events.

Despite these limitations, this study demonstrates the discrepant nature of toxicity identification with a diverse set of notes in a setting which should optimize IRR.

Conclusions
In conclusion, retrospective clinician CTCAE symptom identification and, more broadly, retrospective chart review, are variable and can impact clinical study reporting, generalizability, and intra-study comparison. Standardization efforts and simplification should be considered in study design and future coding terminologies, respectively.

References

1. Trotti A, Colevas AD, Setser A, Basch E. Patient-Reported Outcomes and the Evolution of Adverse Event Reporting in Oncology. *J Clin Oncol*. 2007;25(32):5121-5127. doi:10.1200/JCO.2007.12.4784

2. Atkinson TM, Li Y, Coffey CW, et al. Reliability of adverse symptom event reporting by clinicians. *Qual Life Res*. 2012;21(7):1159-1164. doi:10.1007/s11136-011-0031-4

3. Miller TP, Fisher BT, Getz KD, et al. Unintended consequences of evolution of the Common Terminology Criteria for Adverse Events. *Pediatr Blood Cancer*. 2019;66(7):e27747. doi:10.1002/pbc.27747

4. Miller TP, Li Y, Kavcic M, et al. Accuracy of Adverse Event Ascertainment in Clinical Trials for Pediatric Acute Myeloid Leukemia. *J Clin Oncol*. 2016;34(13):1537-1543. doi:10.1200/JCO.2015.65.5860

5. Revelle W. *Psych: Procedures for Psychological, Psychometric, and Personality Research*.; 2020. https://CRAN.R-project.org/package=psych. Accessed April 13, 2020.

6. Gamer M, Lemon J, Singh IFP. *Irr: Various Coefficients of Interrater Reliability and Agreement*.; 2019. https://CRAN.R-project.org/package=irr. Accessed April 13, 2020.

7. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Medica*. 2012;22(3):276-282.

8. Thanarajasingam G, Hubbard JM, Sloan JA, Grothey A. The Imperative for a New Approach to Toxicity Analysis in Oncology Clinical Trials. *J Natl Cancer Inst*. 2015;107(10):djv216. doi:10.1093/jnci/djv216

9. Atkinson TM, Dueck AC, Satele DV, et al. Clinician vs Patient Reporting of Baseline and Postbaseline Symptoms for Adverse Event Assessment in Cancer Clinical Trials. *JAMA Oncol*. December 2019. doi:10.1001/jamaoncol.2019.5566

10. Hong JC, Tanksley J, Niedzwiecki D, Palta M, Tenenbaum JD. Accuracy of a Natural Language Processing Pipeline to Identify Patient Symptoms during Radiation Therapy. *Int J Radiat Oncol*. 2019;105(1):S70. doi:10.1016/j.ijrobp.2019.06.522

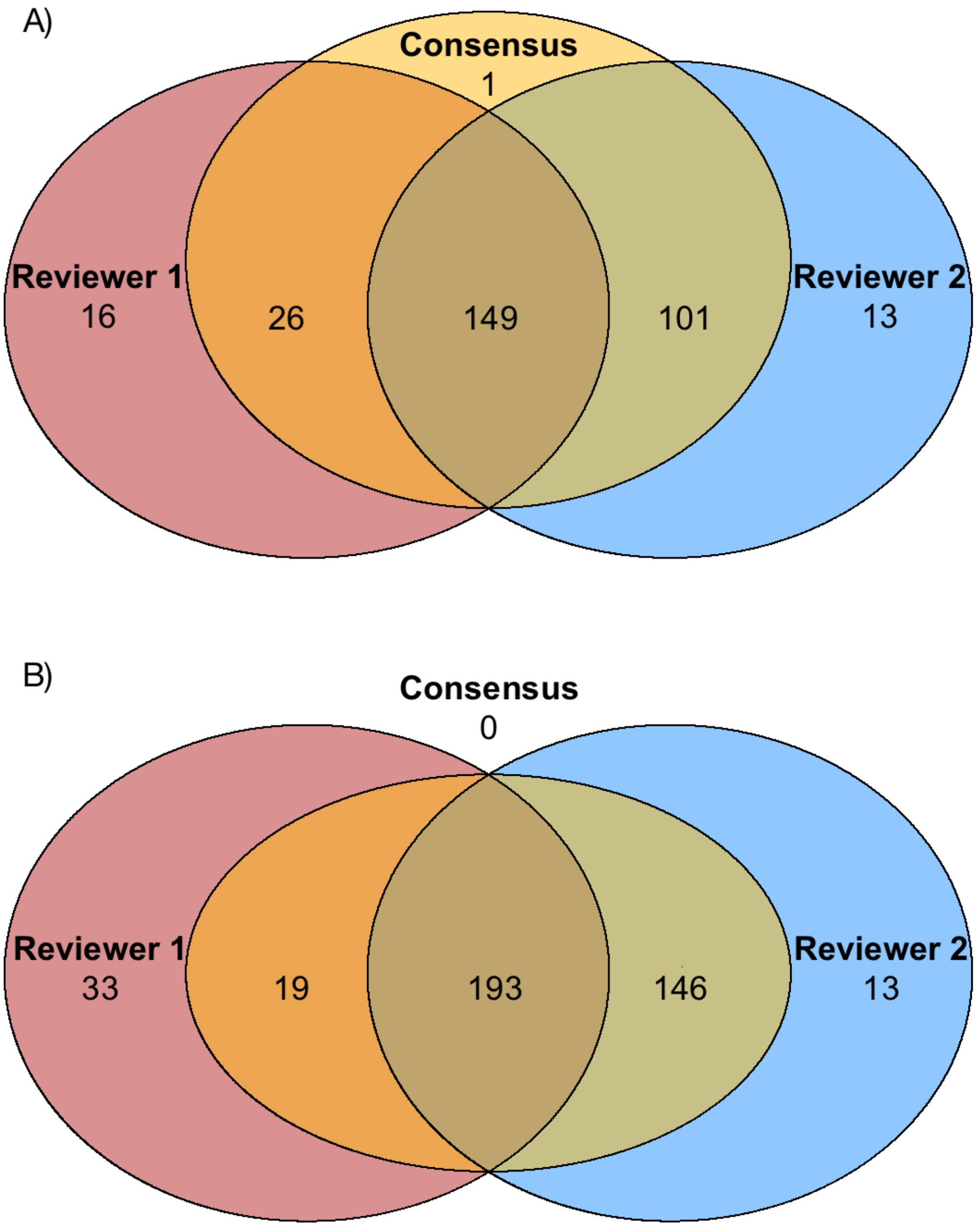

Figure 1. Interrater variability of note-terms with harmonization of terms

With harmonization based on a thesaurus defining synonymous terms, there remained high rates of discordance between reviewers identifying explicitly present (A) and absent (B) symptoms in on-treatment visit notes. Overall weighted kappa coefficient was weighted kappa 0.59 (0.22-1.00). Variability was identified in both noted present and absent symptoms with kappa of 0.6 (0.66-0.71) and 0.6 (0.65-0.69), respectively.

Table 1. Note characteristics and extracted symptoms

Word Count	**Median 203**	**IQR 164.5-237.5**	
Character Count	**Median 1324.5**	**IQR 1103.25-1592.5**	
Number of note authors	**15**		
Disease Site	**Number (N=100)**		
Breast	32		
Head and Neck	15		
Prostate	13		
Central Nervous System	10		
Lung	8		
Gynecologic	7		
Bladder	4		
Metastases (spine, spine, adrenal, leg/lung)	4		
Sarcoma	3		
Esophagus	1		
Skin	1		
Pelvic Lymphoma	1		
Multiple Myeloma	1		
Most common present symptoms	**Number Present (Consensus) (N=100)**	**Reviewers agree present**	**Reviewer discordance**
Dermatitis-Radiation	35	21	15
Fatigue	34	15	20
Pain	24	15	10
Nausea	13	9	4
Pruritus	11	7	4
Cystitis, noninfectious	9	0	**9**
Diarrhea	8	7	1
Urinary Urgency	8	5	2
Mucositis	8	7	3
Folliculitis	7	0	7
Hot Flashes	7	5	2
Total	**277**	**149**	**156**
Most common absent symptoms	**Number Absent (Consensus) (N=100)**	**Reviewers agree absent**	**Reviewer discordance**
Dermatitis-Radiation	42	27	23
Pain	27	18	14
Superficial Soft Tissue Fibrosis	19	0	19
Diarrhea	18	16	2
Seroma	18	16	2
Thrush	16	1	15
Hematochezia	16	15	2
Hematuria	16	13	3
Dysuria	15	12	4
Pruritis	13	6	8
Urinary incontinence	13	12	1
Total	**358**	**193**	**211**

IQR: Interquartile range

*Number present or absent based on consensus adjudication of identifications by both reviewers, rather than the total number of times symptoms were identified by either reviewer.