PLOS BIOLOGY

COMMUNITY PAGE

Ten tips for a text-mining-ready article: How to improve automated discoverability and interpretability

Robert Leaman, Chih-Hsuan Wei, Alexis Allot, Zhiyong Lu*

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, Maryland, United States of America

* zhiyong.lu@nih.gov

Abstract

Data-driven research in biomedical science requires structured, computable data. Increasingly, these data are created with support from automated text mining. Text-mining tools have rapidly matured: although not perfect, they now frequently provide outstanding results. We describe 10 straightforward writing tips—and a web tool, PubReCheck—guiding authors to help address the most common cases that remain difficult for text-mining tools. We anticipate these guides will help authors' work be found more readily and used more widely, ultimately increasing the impact of their work and the overall benefit to both authors and readers. PubReCheck is available at http://www.ncbi.nlm.nih.gov/research/pubrecheck.

OPEN ACCESS

Citation: Leaman R, Wei C-H, Allot A, Lu Z (2020) Ten tips for a text-mining-ready article: How to improve automated discoverability and interpretability. PLoS Biol 18(6): e3000716. https://doi.org/10.1371/journal.pbio.3000716

Published: June 1, 2020

Funding: This research was supported by the NIH Intramural Research Program, National Library of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Provenance: Not commissioned; externally peer reviewed

Introduction

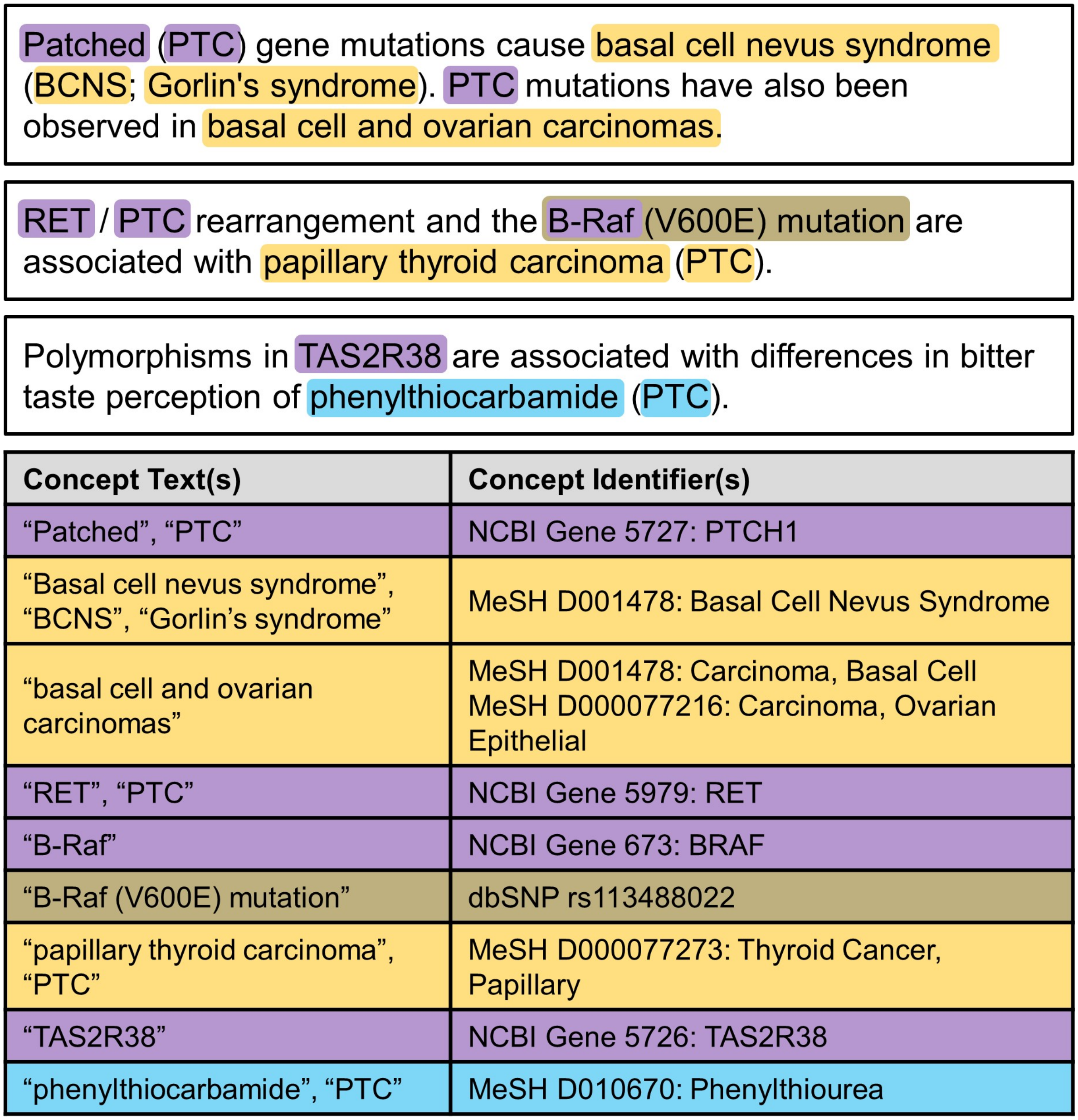
Automated text analysis has proven very effective for helping researchers search the biomedical literature to retrieve relevant articles [1,2]. But as biomedical research becomes increasingly quantitative, the requirement for data-driven research is pushing a need for more specific knowledge embedded within individual articles and for more comprehensive results across the literature [3]. Biocuration addresses these needs by manually extracting the unstructured information in free text articles into structured and computable data in knowledge bases [4]. These curated resources enable connections between seemingly disparate studies and have become essential to current biomedical research. However, data curation at scale remains challenging because of the requirement for significant manual effort by humans [5,6]. Text mining can greatly complement human efforts by automating the conversion of unstructured text such as scientific publications into structured, computable formats, thereby enabling more rapid analyses at a larger scale [7]. Successful uses of text mining in biology include literature-based knowledge discovery [8–11], facilitating analysis of high-throughput (gene expression/genome-wide association) data [12,13], detecting false and contradictory findings [14], and pharmacovigilance [15], among many others.

A critical step in almost all text-mining systems is identifying words and phrases within the text that refer to biomedical concepts. This long-standing task in biomedical text mining was first considered in the 1990s [16] and has been addressed at a number of community-wide challenges since 2004 [17]. The extracted text is then linked with concepts from the relevant biological databases or controlled vocabularies, making the content more accessible, especially to large-scale computational analysis. Fig 1 illustrates concept extraction using extracts from three PubMed articles. Concept recognition systems have matured significantly, identifying a variety of biomedical concepts [18] with performance approaching that of an individual human annotator [19]. Despite significant progress, however, the accuracy of text-mining results remains imperfect.

For text mining to realize its full potential as a powerful method for "seeking a new biology" [23], it is imperative that the critical step of concept extraction be performed as accurately as possible. As illustrated in Fig 1, concept extraction is difficult because of variation and ambiguity, both of which are present to some degree in any natural language. Variation allows concepts to be referenced in multiple ways; for example, Fig 1 shows that the human gene "PTCH1" can also be called "patched" or "PTC," the condition "basal call nevus syndrome" is also known as "Gorlin's syndrome," and "phenylthiourea" and "phenylthiocarbamide" are synonyms. Ambiguity, on the other hand, allows a single phrase to refer to multiple concepts; for example, the acronym "PTC" might refer either to the PTCH1 or RET genes, to papillary thyroid carcinoma, or to phenylthiourea. Managing variation and ambiguity is an important goal of terminology standardization efforts [24,25]. Although we strongly support standardization, we also recognize that authors may not find it practical to identify and apply all relevant standards.

The imperfection of existing automated methods and the difficulty of standardizing terminology has motivated active discussion of alternatives. One proposal suggests that the authors themselves should identify the biomedical concepts referenced in their article [26]. Under this proposal, authors would use controlled vocabularies or ontologies to identify the concepts mentioned in their article prior to publication, similar to the requirement to submit new genes to a central database prior to publication of the manuscript [23]. Unfortunately, this approach requires authors to become knowledgeable in terminologies and curation, which may not be practical. Another proposal is crowdsourcing, in which tasks are outsourced and distributed to many nonexpert workers online, typically by decomposing large tasks into individual decisions [27]. Despite progress, neither alternative has succeeded to date at a large scale.

Meanwhile, recent advances in natural language processing and machine learning continue to improve automated text-mining systems, allowing them to reach an overall accuracy of approximately 80% or higher in many cases. Performance at this level provides mostly outstanding results, with additional help typically only required for the few cases that remain difficult. In this work, we propose 10 writing tips based on a comprehensive analysis of the most prevalent errors experienced by current approaches for automatic concept extraction. We have attempted to make these tips straightforward for authors to implement—regardless of their preferred English dialect—and believe that the suggestions should typically also make the text clearer to human readers. We also provide a web-based tool, PubReCheck (http://www.ncbi.nlm.nih.gov/research/pubrecheck), to help authors visualize the information automated concept extraction tools derive from their text and to automatically identify many types of issues prior to publication. Published research that follows these guides will typically be processed more accurately by automated text analysis tools. We anticipate these guides will allow the author's work to be found more readily and used more widely, ultimately helping the millions of readers who search the biomedical literature satisfy their information needs.



Patched (PTC) gene mutations cause basal cell nevus syndrome (BCNS; Gorlin's syndrome). PTC mutations have also been observed in basal cell and ovarian carcinomas.

RET / PTC rearrangement and the B-Raf (V600E) mutation are associated with papillary thyroid carcinoma (PTC).

Polymorphisms in TAS2R38 are associated with differences in bitter taste perception of phenylthiocarbamide (PTC).

Concept Text(s)	Concept Identifier(s)
“Patched”, “PTC”	NCBI Gene 5727: PTCH1
“Basal cell nevus syndrome”, “BCNS”, “Gorlin’s syndrome”	MeSH D001478: Basal Cell Nevus Syndrome
“basal cell and ovarian carcinomas”	MeSH D001478: Carcinoma, Basal Cell MeSH D000077216: Carcinoma, Ovarian Epithelial
“RET”, “PTC”	NCBI Gene 5979: RET
“B-Raf”	NCBI Gene 673: BRAF
“B-Raf (V600E) mutation”	dbSNP rs113488022
“papillary thyroid carcinoma”, “PTC”	MeSH D000077273: Thyroid Cancer, Papillary
“TAS2R38”	NCBI Gene 5726: TAS2R38
“phenylthiocarbamide”, “PTC”	MeSH D010670: Phenylthiourea

Fig 1. Concept extraction is a critical step in text mining. Phrases referring to a concept are associated with an identifier from an appropriate database (purple = gene, orange = disease, brown = mutation = brown, cyan = chemical). Text-mining systems must handle variation—“patched” and “PTC” both refer to “PTCH1”—and ambiguity—“PTC” could refer to “PTCH1,” “RET,” “papillary thyroid carcinoma,” or “phenylthiourea.” Examples adapted from [20–22].

https://doi.org/10.1371/journal.pbio.3000716.g001

Top 10 Tips

Tip 1: Clearly associate gene and protein names with the species

Entries in gene and protein databases are differentiated by species: in humans, the BRCA1 gene is NCBI Gene 672, but in mice (*Mus musculus*) it is NCBI Gene 12189. Automated systems for identifying genes and proteins must, therefore, determine the species first. If the species is not mentioned directly, the system must try to infer it from related concepts: the cell line “GH(3)” implies *Rattus norvegicus* [28] and the strain “TA100” signifies *Salmonella* *typhimurium* [29]. Although inferring the species from context is often effective, directly stating the species significantly reduces the potential for error. We recommend authors clearly mention the relevant species when discussing genes or proteins whenever possible. This is especially important the first time it is mentioned: for example, "we investigated the role of Brca1 during mouse embryonic cortical development . . ." [30]. This recommendation does not apply, however, in cases when the discussion is not with respect to any specific species, such as when referring to (homologous) genes in a general context.

Tip 2: Supply context critical for comprehension prominently and in proximity

Like human readers, text-mining systems use the surrounding context to help resolve ambiguous words and phrases. For example, whereas "p24" could refer to at least four different human proteins, the text "p24 also maps to chromosome 12" [31] must refer to the product of human gene TMED2 (NCBI Gene 10959), because the other possibilities are assigned to other chromosomes. Synonyms are especially useful for clarifying the intended meaning; for example, the synonym provided in the text, "which interacts with human NAP1 (NCKAP1) . . ." [32], narrows the number of possibilities for "NAP1" from 9 to 1 (NCBI Gene 10787). The need for clarifying context may arise because a specific name is ambiguous, or it may arise because a specific relationship is required, such as the species for a gene or a gene name for a genetic variant.

Although context can resolve ambiguity effectively, it is only helpful if the correct association can be identified. This is especially important for the abstract: it summarizes the article, is often provided without the full article text, and is the focus of many text-mining methods—though work on mining full-text articles is well underway [33–35]. Current automated methods for context association are more accurate when the related information is within the same sentence or, at most, within the same paragraph. Although the abstract cannot contain the details necessary to analyze its claims, if comprehending the summary provided by the abstract requires context from other sections—for example, to determine which "p24" protein is the subject of the article—then the risk of error is significantly increased. We recommend authors provide context that is critical for comprehension (such as identifying ambiguous concept names) prominently—in the abstract—and in proximity, preferably in the same sentence.

Tip 3: Define abbreviations and acronyms

Abbreviations and acronyms allow cumbersome terms to be referenced more concisely: "acute myeloid leukemia" could become "AML." Although a few carefully chosen acronyms can improve readability, acronyms can also be ambiguous because they typically have more than one possible meaning. For example, "AML" often refers to "acute myeloid leukemia," but it is also used for "angiomyolipoma," "anterior mitral leaflet," "amlodipine," "amoxicillin," and "amiloride." The intended meaning of an acronym is therefore usually provided at first use, as in "Patients with acute myeloid leukemia (AML) are . . ." [36]. Without such a definition, a human reader or text-mining system must infer the meaning of an acronym from context, which is often unclear. We recommend all abbreviations and acronyms be listed with the corresponding full term the first time they are used.

Tip 4: Refer to concepts by name

Language is variable: it can communicate ideas in multiple ways. Accordingly, a text might refer to a concept by name ("orthostatic hypotension") or with a description ("immediate drop in systolic blood pressure observed on standing" [37]). Descriptions can be very helpful, but names provide several important advantages for automated tools. Names are usually easier for automated tools to locate because they have a simpler structure and less freedom for variation than descriptions. This makes them easier to differentiate from the surrounding text and match against the names in a controlled vocabulary, thus identifying them as referring to the corresponding concept. Names are also shorter than descriptions and thus have fewer places where the reference could potentially begin or end, which translates to fewer opportunities for error. We recommend referring to concepts primarily by name; when a description is needed, we suggest providing both.

Tip 5: Use one term per concept

Synonyms can help authors clarify their intended meaning when a concept is first introduced, as in "The mouse Foxq1 gene, also known as Hfh1 . . ." [38] or "Primary hyperaldosteronism (Conn's syndrome) is . . ." [39]. However, using multiple terms interchangeably without clearly indicating that they should be considered equivalent is confusing at best and at worst may cause the reader—or text-mining system—to consider them to refer to different concepts. For automated algorithms, even minor variations such as hyphenation ("gastroenteritis" vs "gastro-enteritis" [40]) or possessives ("Schiff bases" vs "Schiff's bases" [41]) require additional handling, increasing the risk of error. We recommend authors choose a term for each concept and use it consistently—varying from the exact text chosen only when required, such as for capitalization or plurals.

Tip 6: Prefer the complete and precise term

Some biomedical terms have multiple meanings and are therefore ambiguous, despite being commonly used in the literature. The term "yeast" often refers specifically to the model organism *Saccharomyces cerevisiae*, but there are over 1,500 known species of yeast—one of which (*Candida auris*) is an emerging global health threat [42]. Similarly, "mouse" frequently refers to *M. musculus*, but "mouse" could refer to any species in the genus or to the genus itself. Subtypes also matter: despite some shared characteristics, "type 1 diabetes" and "type 2 diabetes" have many clinically relevant differences. Other ambiguous terms involve a close relationship between concepts. For example, "Epstein-Barr" probably refers to either to "Epstein-Barr virus" or "Epstein-Barr infection" and "Multiple Endocrine Neoplasia Type 1 (MEN1)" refers to "MEN1 syndrome" or "MEN1 gene." We recommend preferring the precise and complete scientific term wherever possible. If the full term is too cumbersome, we suggest clarifying a more convenient term at first use, as in "The laboratory rat (Rattus norvegicus) is . . ." [43]. However, detailed classifications that are irrelevant or uncertain should be withheld, as in "Danio sp. could therefore play a significant role in controlling mosquito breeding . . ." [44].

Tip 7: Coordinate compound terms cautiously

Compound terms such as "pineal tumour" and "retinal tumour" are often combined or coordinated to form a single phrase, as in "pineal and retinal tumours" [45]. Although simple coordinated phrases often improve readability, the number of possible interpretations for complex coordinated phrases quickly make them difficult to interpret. For example, a human reader might be able to determine that the phrase "unstimulated and MTb- or LPS-stimulated THP-1 cells" [46] refers to "unstimulated THP-1 cells," "MTb-stimulated THP-1 cells," and "LPS-stimulated THP-1 cells." However, an automated tool must consider a vastly greater number of possibilities—such as "unstimulated LPS-stimulated THP-1 cells" and "MTb-1 cells"—and discard them as nonsensical. The risk that an automated tool will incorrectly prefer the wrong interpretation therefore depends strongly on the complexity of the coordinated phrase [47,48].

Simplifying the phrase—for example, to “unstimulated, MTb-stimulated, and LPS-stimulated THP-1 cells”—greatly reduces the complexity and thus the opportunities for error. We recommend that authors avoid creating coordinated compound terms with multiple potential interpretations.

Tip 8: Spacing matters

Automated text-mining algorithms often initially identify meaningful units of text (such as sentences and words) and then proceed to interpret each unit. This approach is efficient and generally effective but tends to encounter problems if boundary markers such as spaces or periods are misplaced or missing. For example, the space missing from “ipratropiumbromide” [49] will make it more difficult to automatically recognize that it refers to the drug “ipratropium bromide.” Similarly, the space missing in “BRAFV600E” [50] will increase the difficulty of identifying the gene "BRAF" and the mutation "V600E." Extraneous spaces (“malignant lym phoma,” [51]) cause similar issues. We recommend identifying spacing issues with careful proofreading and spell-checking.

Tip 9: Verify parentheses and brackets are correctly paired

Missing parentheses are also a concern, as in “ultrafiltration during cardiopulmonary bypass ICPB) has mandated . . .” [52]. Because parenthetical text is typically used to provide readers with information outside of the main narrative [53], correctly interpreting text with mismatching parentheses is difficult. Moreover, it is not straightforward for automated systems to determine whether a parenthesis should be added (and if so, which location) or removed (and if so, which one). We recommend verifying that all parentheses are correctly paired and no parentheses are placed extraneously.

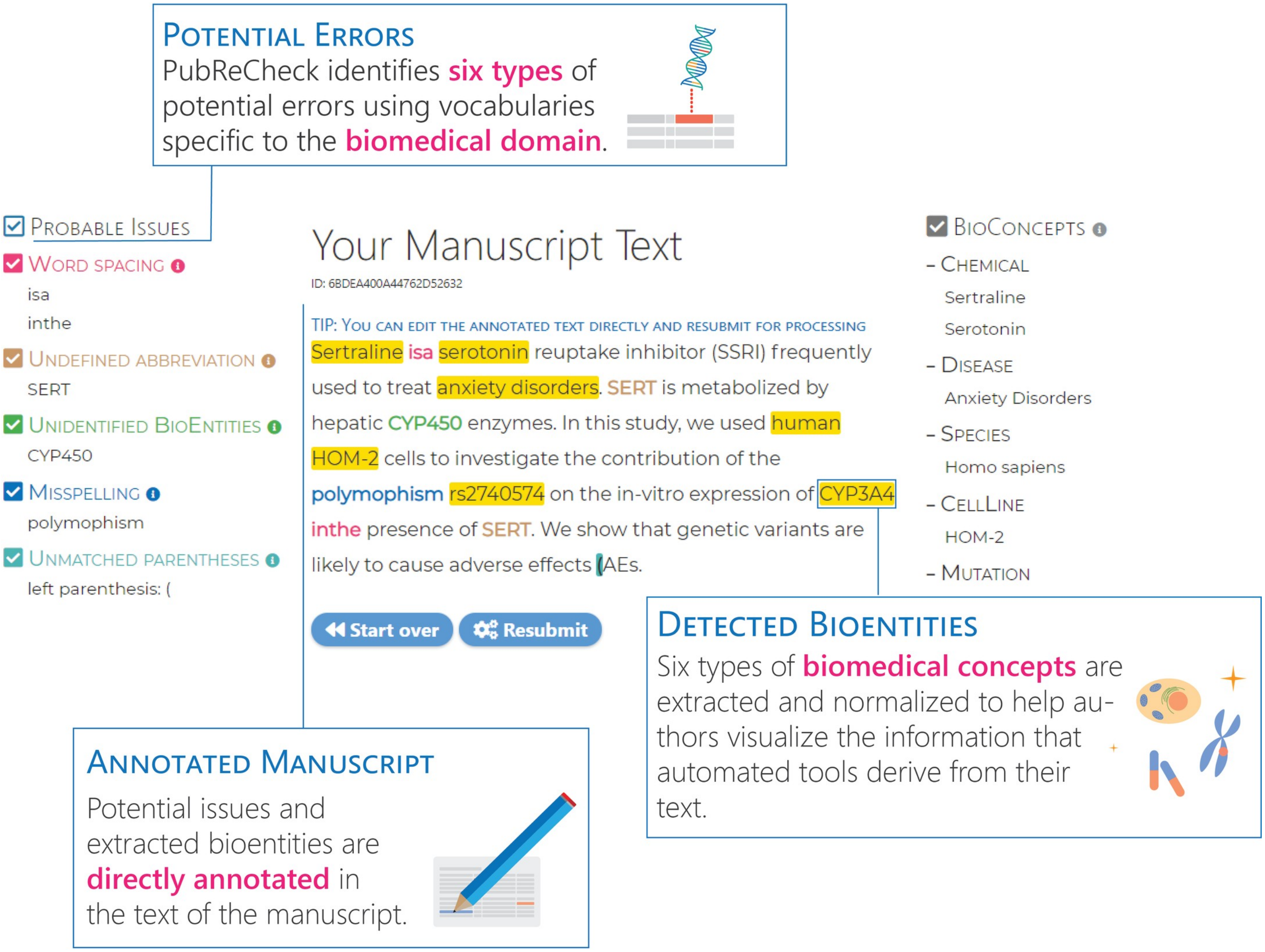
Tip 10: Recheck spelling with a different method

Biomedical terminology can be difficult to spell correctly, and it is not hard to find misspellings in the published literature. For example, not only does PubMed contain many examples of “hemorrhage” misspelled as “hemmorhage,” it also contains the plural (“hemmorhages”) and at least two related forms (“autohemmorhage” and “microhemmorhage”). Although automated tools can help with misspellings, the need to balance correcting errors against the possibility of introducing new ones makes automatic spelling correction more difficult than suggesting potential corrections to the user. On the other hand, recent text-mining tools should be able to handle true spelling variations (e.g., “leukemia” versus “leukaemia”).

Although careful proofreading is always valuable, proofreading manually may be subject to diminishing returns. We recommend rechecking for misspellings using a method not used previously. For example, authors could request assistance from a colleague not previously involved or use a different spell-checking tool. If nothing else, we suggest carefully rechecking each word in the title and abstract for any remaining spelling errors.

Automated Tool: PubReCheck

To help authors automatically identify many types of issues prior to publication, we developed a web-based tool, PubReCheck (http://www.ncbi.nlm.nih.gov/research/pubrecheck). PubReCheck provides two primary functions, as shown in the screenshot of its results on a synthetic abstract in Fig 2. First, PubReCheck identifies six types of biomedical concepts: genes, diseases, chemicals, genetic variants, species, and cell lines. These results help authors visualize the information that automated tools derive from their text, allowing the text to be rephrased if



POTENTIAL ERRORS

PubReCheck identifies **six types** of potential errors using vocabularies specific to the **biomedical domain**.

☑ PROBABLE ISSUES

☑ WORD SPACING

isa

inthe

☑ UNDEFINED ABBREVIATION

SERT

☑ UNIDENTIFIED BIOENTITIES

CYP450

☑ MISSPELLING

polymophism

☑ UNMATCHED PARENTHESES

left parenthesis: (

Your Manuscript Text

ID: 6BDEA400A44762D52632

TIP: YOU CAN EDIT THE ANNOTATED TEXT DIRECTLY AND RESUBMIT FOR PROCESSING

Sertraline isa serotonin reuptake inhibitor (SSRI) frequently used to treat anxiety disorders. SERT is metabolized by hepatic CYP450 enzymes. In this study, we used human HOM-2 cells to investigate the contribution of the polymophism rs2740574 on the in-vitro expression of CYP3A4 inthe presence of SERT. We show that genetic variants are likely to cause adverse effects (AEs.

☑ BIOCONCEPTS

- CHEMICAL

Sertraline

Serotonin

- DISEASE

Anxiety Disorders

- SPECIES

Homo sapiens

- CELLLINE

HOM-2

- MUTATION

DETECTED BIOENTITIES

Six types of **biomedical concepts** are extracted and normalized to help authors visualize the information that automated tools derive from their text.

ANNOTATED MANUSCRIPT

Potential issues and extracted bioentities are **directly annotated** in the text of the manuscript.

Fig 2. Screenshot of the PubReCheck system using an artificial abstract. PubReCheck identifies six types of biomedical concepts: genes, diseases, chemicals, genetic variants, species, and cell lines. PubReCheck also identifies six types of potential issues: misspellings, word spacing errors, undefined abbreviations, entities that cannot be uniquely identified, novel words, and unmatched parentheses.

https://doi.org/10.1371/journal.pbio.3000716.g002

required. PubReCheck also directly identifies six types of potential errors: misspellings, word spacing errors, novel words, undefined abbreviations, entities that cannot be uniquely identified, and unmatched parentheses. Like many automated tools, PubReCheck identifies misspellings (Tip 10) but is adapted to handle biomedical vocabulary that is rare in the general domain. Similarly, PubReCheck identifies potential word spacing errors (Tip 8)—phrases that need a space added or removed—and words that are uncommon in the biomedical literature, even if they may not be marked as a misspelling. PubReCheck also identifies undefined abbreviations (Tip 3) and unmatched parentheses and brackets (Tip 9). Finally, PubReCheck locates biomedical concepts that cannot be uniquely identified, which addresses several issues that may result in an ambiguous concept name (Tips 1, 2, and 6, primarily).

Automatically identifying and correcting potential errors is itself subject to errors. Error-checking the existing biomedical literature directly is problematic because it is difficult to recover from the additional errors that may be introduced. PubReCheck instead provides feedback directly to authors, who have both a strong interest in ensuring no errors remain and the ability to simply ignore a few phrases incorrectly identified as containing an error. PubReCheck therefore intentionally prioritizes identifying more potential errors. Moreover, because PubReCheck also allows authors to correct issues prior to publication, the maximum benefit is provided to both authors and readers.

Conclusion

The continued rapid expansion of the biomedical literature necessitates the use of automated methods to address the information overload. Moreover, the increase in quantitative research in biology motivates moving beyond retrieving articles to extracting and converting their content to structured formats that enable computational processing. Although the accuracy of text-mining methods has improved dramatically in recent years—and will likely continue to improve—several issues remain difficult to address automatically.

Complementary to calls and initiatives that ask authors to follow standards and use standardized terminology, we have proposed a set of straightforward writing tips, summarized in Box 1, to help authors provide the information necessary to help automated text-mining algorithms to process their articles correctly. These tips—and especially our online tool, PubReCheck—may also be useful for editors, reviewers, publishers, and proofreaders. Although additional suggestions (and exemptions) could be identified, the ones presented have been chosen as likely to provide significant benefit for relatively modest effort. Articles that follow these tips will typically be processed more accurately, allowing their content to be found more readily and used more widely, thereby increasing its impact. Following these guidelines at a large scale will improve the ability of individual researchers to find the articles that meet their information needs. In short, following these tips will help us help you, and millions.

Box 1. Summary of our recommendations to help articles be processed more accurately

- Clearly mention the relevant species when discussing genes or proteins
- Supply context critical for comprehension prominently and in proximity
- Define abbreviations and acronyms the first time they are used
- Refer to concepts primarily by name, not description
- Choose a term for each concept and use it consistently
- Prefer the complete and precise scientific term
- Avoid creating complex coordinated compound terms
- Recheck for word spacing errors
- Verify parentheses and brackets are correctly paired
- Recheck for misspellings using a different method

Acknowledgments

We thank Dr. John Wilbur for helpful discussions of this work.

References

1. Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, et al. Best Match: New relevance search for PubMed. PLoS Biol. 2018; 16(8):e2005343. https://doi.org/10.1371/journal.pbio.2005343 PMID: 30153250
2. Fiorini N, Leaman R, Lipman DJ, Lu Z. How user intelligence is improving PubMed. Nat Biotechnol. Epub 2018 Oct 1.
3. Markowetz F. All biology is computational biology. PLoS Biol. 2017; 15(3):e2002050. https://doi.org/10.1371/journal.pbio.2002050 PMID: 28278152
4. International Society for Biocuration. Biocuration: Distilling data into knowledge. PLoS Biol. 2018; 16(4): e2002846. https://doi.org/10.1371/journal.pbio.2002846 PMID: 29659566
5. Baumgartner WA Jr., Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. Bioinformatics. 2007; 23(13):i41–8. https://doi.org/10.1093/bioinformatics/btm229 PMID: 17646325
6. Bourne PE, Lorsch JR, Green ED. Perspective: Sustaining the big-data ecosystem. Nature. 2015; 527 (7576):S16–7. https://doi.org/10.1038/527S16a PMID: 26536219
7. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006; 7(2):119–29. https://doi.org/10.1038/nrg1768 PMID: 16418747
8. Choi BK, Dayaram T, Parikh N, Wilkins AD, Nagarajan M, Novikov IB, et al. Literature-based automated discovery of tumor suppressor p53 phosphorylation and inhibition by NEK2. Proc Natl Acad Sci U S A. 2018; 115(42):10666–71. https://doi.org/10.1073/pnas.1806643115 PMID: 30266789
9. Gyori BM, Bachman JA, Subramanian K, Muhlich JL, Galescu L, Sorger PK. From word models to executable models of signaling networks using automated assembly. Mol Syst Biol. 2017; 13(11):954. https://doi.org/10.15252/msb.20177651 PMID: 29175850
10. Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods. 2019; 16(6):505–7. https://doi.org/10.1038/s41592-019-0422-y PMID: 31110280
11. Perez-Iratxeta C, Bork P, Andrade MA. Association of genes to genetically inherited diseases using data mining. Nat Genet. 2002; 31(3):316–9. https://doi.org/10.1038/ng895 PMID: 12006977
12. Natarajan J, Berrar D, Dubitzky W, Hack C, Zhang Y, DeSesa C, et al. Text mining of full-text journal articles combined with gene expression analysis reveals a relationship between sphingosine-1-phosphate and invasiveness of a glioblastoma cell line. BMC Bioinformatics. 2006; 7:373. https://doi.org/10.1186/1471-2105-7-373 PMID: 16901352
13. Huang LC, Ross KE, Baffi TR, Drabkin H, Kochut KJ, Ruan Z, et al. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. Sci Rep. 2018; 8(1):6518. https://doi.org/10.1038/s41598-018-24457-1 PMID: 29695735
14. Rzhetsky A, Iossifov I, Loh JM, White KP. Microparadigms: chains of collective reasoning in publications about molecular interactions. Proc Natl Acad Sci U S A. 2006; 103(13):4940–5. https://doi.org/10.1073/pnas.0600591103 PMID: 16543380
15. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S, et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform. 2015; 54:202–12. https://doi.org/10.1016/j.jbi.2015.02.004 PMID: 25720841
16. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput. 1998:707–18. PMID: 9697224
17. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2016; 17(1):132–44. https://doi.org/10.1093/bib/bbv024 PMID: 25935162
18. Wei CH, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019; 47(W1):W587–W93. https://doi.org/10.1093/nar/gkz389 PMID: 31114887
19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020; 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682 PMID: 31501885

20. Zedan W, Robinson PA, High AS. A novel polymorphism in the PTC gene allows easy identification of allelic loss in basal cell nevus syndrome lesions. Diagn Mol Pathol. 2001; 10(1):41–5. https://doi.org/10.1097/00019606-200103000-00007 PMID: 11277394

21. Caria P, Dettori T, Frau DV, Borghero A, Cappai A, Riola A, et al. Assessing RET/PTC in thyroid nodule fine-needle aspirates: the FISH point of view. Endocr Relat Cancer. 2013; 20(4):527–36. https://doi.org/10.1530/ERC-13-0157 PMID: 23722226

22. Wooding S, Bufe B, Grassi C, Howard MT, Stone AC, Vazquez M, et al. Independent evolution of bitter-taste sensitivity in humans and chimpanzees. Nature. 2006; 440(7086):930–4. https://doi.org/10.1038/nature04655 PMID: 16612383

23. Rzhetsky A, Seringhaus M, Gerstein M. Seeking a new biology through text mining. Cell. 2008; 134 (1):9–13. https://doi.org/10.1016/j.cell.2008.06.029 PMID: 18614002

24. Chute CG, Cohn SP, Campbell JR. A framework for comprehensive health terminology systems in the United States: development guidelines, criteria for selection, and public policy implications. ANSI Healthcare Informatics Standards Board Vocabulary Working Group and the Computer-Based Patient Records Institute Working Group on Codes and Structures. J Am Med Inform Assoc. 1998; 5(6):503–10. https://doi.org/10.1136/jamia.1998.0050503 PMID: 9824798

25. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998; 37(4–5):394–403. PMID: 9865037

26. Leitner F, Chatr-aryamontri A, Mardis SA, Ceol A, Krallinger M, Licata L, et al. The FEBS Letters/BioCreative II.5 experiment: making biological information accessible. Nat Biotechnol. 2010; 28(9):897–9. https://doi.org/10.1038/nbt0910-897 PMID: 20829821

27. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform. 2016; 17(1):23–32. https://doi.org/10.1093/bib/bbv021 PMID: 25888696

28. Tsai HW, Katzenellenbogen JA, Katzenellenbogen BS, Shupnik MA. Protein kinase A activation of estrogen receptor alpha transcription does not require proteasome activity and protects the receptor from ligand-mediated degradation. Endocrinology. 2004; 145(6):2730–8. https://doi.org/10.1210/en.2003-1470 PMID: 15033909

29. Petta TB, de Medeiros SR, do Egito ES, Agnez-Lima LF. Genotoxicity induced by saponified coconut oil surfactant in prokaryote systems. Mutagenesis. 2004; 19(6):441–4. https://doi.org/10.1093/mutage/geh054 PMID: 15548754

30. Pulvers JN, Huttner WB. Brca1 is required for embryonic development of the mouse cerebral cortex to normal size by preventing apoptosis of early neural progenitors. Development. 2009; 136(11):1859–68. https://doi.org/10.1242/dev.033498 PMID: 19403657

31. Katz FE, Parkar M, Stanley K, Murray LJ, Clark EA, Greaves MF. Chromosome mapping of cell membrane antigens expressed on activated B cells. Eur J Immunol. 1985; 15(1):103–6. https://doi.org/10.1002/eji.1830150121 PMID: 3871395

32. Yamamoto A, Suzuki T, Sakaki Y. Isolation of hNap1BP which interacts with human Nap1 (NCKAP1) whose expression is down-regulated in Alzheimer's disease. Gene. 2001; 271(2):159–69. https://doi.org/10.1016/s0378-1119(01)00521-2 PMID: 11418237

33. Westergaard D, Staerfeldt HH, Tonsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018; 14(2):e1005962. https://doi.org/10.1371/journal.pcbi.1005962 PMID: 29447159

34. Allot A, Peng Y, Wei CH, Lee K, Phan L, Lu Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Res. 2018; 46(W1):W530–W6. https://doi.org/10.1093/nar/gky355 PMID: 29762787

35. Allot A, Chen Q, Kim S, Vera Alvarez R, Comeau DC, Wilbur WJ, et al. LitSense: making sense of biomedical literature at sentence level. Nucleic Acids Res. 2019; 47(W1):W594–W9. https://doi.org/10.1093/nar/gkz289 PMID: 31020319

36. Sierra J, Szer J, Kassis J, Herrmann R, Lazzarino M, Thomas X, et al. A single dose of pegfilgrastim compared with daily filgrastim for supporting neutrophil recovery in patients treated for low-to-intermediate risk acute myeloid leukemia: results from a randomized, double-blind, phase 2 trial. BMC Cancer. 2008; 8:195. https://doi.org/10.1186/1471-2407-8-195 PMID: 18616811

37. Mytton OT, Simpson A, Thompson AA, Oram RA, Darowski A, Yu LM, et al. Manual assessment of the initial fall in blood pressure after orthostatic challenge at high altitude. Wilderness Environ Med. 2008; 19(4):225–32. https://doi.org/10.1580/07-WEME-OR-097.1 PMID: 19099326

38. Goering W, Adham IM, Pasche B, Manner J, Ochs M, Engel W, et al. Impairment of gastric acid secretion and increase of embryonic lethality in Foxq1-deficient mice. Cytogenet Genome Res. 2008; 121 (2):88–95. https://doi.org/10.1159/000125833 PMID: 18544931

39. Schirpenbach C, Segmiller F, Diederich S, Hahner S, Lorenz R, Rump LC, et al. The diagnosis and treatment of primary hyperaldosteronism in Germany: results on 555 patients from the German Conn Registry. Dtsch Arztebl Int. 2009; 106(18):305–11. https://doi.org/10.3238/arztebl.2009.0305 PMID: 19547646

40. de Crom SC, Rossen JW, de Moor RA, Veldkamp EJ, van Furth AM, Obihara CC. Prospective assessment of clinical symptoms associated with enterovirus and parechovirus genotypes in a multicenter study in Dutch children. J Clin Virol. 2016; 77:15–20. https://doi.org/10.1016/j.jcv.2016.01.014 PMID: 26875098

41. Kumar S, Kumar P, Sati N. Synthesis and biological evaluation of Schiff bases and azetidinones of 1-naphthol. J Pharm Bioallied Sci. 2012; 4(3):246–9. https://doi.org/10.4103/0975-7406.99066 PMID: 22923968

42. Spivak ES, Hanson KE. Candida auris: an Emerging Fungal Pathogen. J Clin Microbiol. 2018; 56(2): e01588–17. https://doi.org/10.1128/JCM.01588-17 PMID: 29167291

43. van Boxtel R, Toonen PW, Verheul M, van Roekel HS, Nijman IJ, Guryev V, et al. Improved generation of rat gene knockouts by target-selected mutagenesis in mismatch repair-deficient animals. BMC Genomics. 2008; 9:460. https://doi.org/10.1186/1471-2164-9-460 PMID: 18840264

44. Haq S, Prasad H, Prasad RN, Sharma T. Availability and utility of local fishes of Shahjahanpur for mosquito control. Indian J Malariol. 1993; 30(1):1–8. PMID: 8100538

45. Onadim Z, Woolford AJ, Kingston JE, Hungerford JL. The RB1 gene mutation in a child with ectopic intracranial retinoblastoma. Br J Cancer. 1997; 76(11):1405–9. https://doi.org/10.1038/bjc.1997.570 PMID: 9400934

46. Falvo JV, Tsytsykova AV, Goldfeld AE. Transcriptional control of the TNF gene. Curr Dir Autoimmun. 2010; 11:27–60. https://doi.org/10.1159/000289196 PMID: 20173386

47. Wei CH, Leaman R, Lu Z. SimConcept: a hybrid approach for simplifying composite named entities in biomedical text. IEEE J Biomed Health Inform. 2015; 19(4):1385–91. https://doi.org/10.1109/JBHI.2015.2422651 PMID: 25879978

48. Chae J, Jung Y, Lee T, Jung S, Huh C, Kim G, et al. Identifying non-elliptical entity mentions in a coordinated NP with ellipses. J Biomed Inform. 2014; 47:139–52. https://doi.org/10.1016/j.jbi.2013.10.002 PMID: 24153413

49. Kikis D, Esser H, Heinrich K. Influence of ipratropiumbromide on heart rate and hemodynamics in patients with sinus bradycardia. Clin Cardiol. 1982; 5(8):441–5. https://doi.org/10.1002/clc.4960050804 PMID: 6215204

50. Moon HJ, Kwak JY, Kim EK, Choi JR, Hong SW, Kim MJ, et al. The role of BRAFV600E mutation and ultrasonography for the surgical management of a thyroid nodule suspicious for papillary thyroid carcinoma on cytology. Ann Surg Oncol. 2009; 16(11):3125–31. https://doi.org/10.1245/s10434-009-0644-9 PMID: 19644722

51. Fahey JL, Finegold I, Rabson AS, Manaker RA. Immunoglobulin synthesis in vitro by established human cell lines. Science. 1966; 152(3726):1259–61. https://doi.org/10.1126/science.152.3726.1259 PMID: 5937117

52. Glogowski KR, Stammers AH, Niimi KS, Tremain KD, Muhle ML, Trowbridge CC. The effect of priming techniques of ultrafiltrators on blood rheology: an in vitro evaluation. Perfusion. 2001; 16(3):221–8. https://doi.org/10.1177/026765910101600308 PMID: 11419658

53. Cohen KB, Christiansen T, Hunter LE. Parenthetically speaking: classifying the contents of parentheses for text mining. AMIA Annu Symp Proc. 2011; 2011:267–72. PMID: 22195078