



HAL
open science

Misspellings or "miscellings"-Non-verifiable and unknown cell lines in cancer research publications

Danielle J. Oste, Pranujan Pathmendra, Reese A K Richardson, Gracen Johnson, Yida Ao, Maya D Arya, Naomi R Enochs, Muhammad Hussein, Jinghan Kang, Aaron Lee, et al.

► To cite this version:

Danielle J. Oste, Pranujan Pathmendra, Reese A K Richardson, Gracen Johnson, Yida Ao, et al.. Misspellings or "miscellings"-Non-verifiable and unknown cell lines in cancer research publications. International Journal of Cancer, 2024, pp.1–12. 10.1002/ijc.34995 . hal-04577886

HAL Id: hal-04577886

<https://ut3-toulouseinp.hal.science/hal-04577886>

Submitted on 16 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.




Distributed under a Creative Commons Attribution 4.0 International License

RESEARCH ARTICLE

Innovative Tools and Methods

Misspellings or “miscellings”—Non-verifiable and unknown cell lines in cancer research publications

Danielle J. Oste^{1,2} | Pranuwan Pathmendra¹ | Reese A. K. Richardson³ |
 Gracen Johnson¹ | Yida Ao¹ | Maya D. Arya¹ | Naomi R. Enochs¹ |
 Muhammed Hussein¹ | Jinghan Kang¹ | Aaron Lee¹ | Jonathan J. Danon⁴ |
 Guillaume Cabanac^{5,6} | Cyril Labbé⁷ | Amanda Capes Davis⁸ |
 Thomas Stoeger^{9,10,11} | Jennifer A. Byrne^{1,12} 

¹School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

²Sydney School of Veterinary Science, Faculty of Science, The University of Sydney, Sydney, New South Wales, Australia

³Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois, USA

⁴School of Chemistry, Faculty of Science, The University of Sydney, Sydney, New South Wales, Australia

⁵IRIT UMR 5505 CNRS, University of Toulouse, Toulouse, France

⁶Institut Universitaire de France (IUF), Paris, France

⁷CNRS, Grenoble INP, Laboratoire d'Informatique de Grenoble, Université Grenoble Alpes, Grenoble, France

⁸CellBank Australia, Children's Medical Research Institute, The University of Sydney, Sydney, New South Wales, Australia

⁹Feinberg School of Medicine in the Division of Pulmonary and Critical Care Medicine, Northwestern University, Chicago, Illinois, USA

¹⁰The Potocsnak Longevity Institute, Northwestern University, Chicago, Illinois, USA

¹¹Simpson Querrey Lung Institute for Translational Science, Chicago, Illinois, USA

¹²NSW Health Statewide Biobank, NSW Health Pathology, Sydney, New South Wales, Australia

Correspondence

Jennifer A. Byrne, NSW Health Pathology and University of Sydney, Sydney, NSW, Australia.
 Email: jennifer.byrne@health.nsw.gov.au

Funding information

University of Sydney; NIH(USA), Grant/Award Number: AG068544; Northwestern University; National Health and Medical Research Council of Australia (NHMRC) Ideas Grant, Grant/Award Number: APP1184263; NHMRC Investigator Grant, Grant/Award Number: GNT2008066; Moderna Inc

Abstract

Reproducible laboratory research relies on correctly identified reagents. We have previously described gene research papers with wrongly identified nucleotide sequence(s), including papers studying *miR-145*. Manually verifying reagent identities in 36 recent *miR-145* papers found that 56% and 17% of papers described misidentified nucleotide sequences and cell lines, respectively. We also found 5 cell line identifiers in *miR-145* papers with misidentified nucleotide sequences and cell lines, and 18 cell line identifiers published elsewhere, that did not represent indexed human cell lines. These 23 identifiers were described as non-verifiable (NV), as their identities were unclear. Studying 420 papers that mentioned 8 NV identifier(s) found 235 papers (56%) that referred to 7 identifiers (BGC-803, BSG-803, BSG-823, GSE-1, HGC-7901, HGC-803, and MGC-823) as independent cell lines. We could not find any publications describing how these cell lines were established. Six cell lines were sourced from cell line repositories with externally accessible online catalogs, but these

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *International Journal of Cancer* published by John Wiley & Sons Ltd on behalf of UICC.

cell lines were not indexed as claimed. Some papers also stated that short tandem repeat (STR) profiles had been generated for three cell lines, yet no STR profiles could be identified. In summary, as NV cell lines represent new challenges to research integrity and reproducibility, further investigations are required to clarify their status and identities.

KEYWORDS

cancer, cell lines, non-verifiable, reagents, wrongly identified

What's new

Reproducible laboratory research relies on correctly identified reagents. Here, the authors flagged 23 non-verifiable (NV) human cell line identifiers in recent papers. Although some NV identifiers are likely mere misspellings of known cell lines, the results indicate that some misspelled cell lines can gain new identities as independent cell lines, a process the authors describe as “miscelling.” Of the eight identifiers studied in detail, seven NV identifiers were unexpectedly referred to as independent cell lines across 235 publications, lacking a description of how they were established, not appearing in the claimed external repositories, and having no short tandem repeat profile.

1 | INTRODUCTION

Preclinical research aims to produce reliable observations that can be further tested or extended through translational studies. A variety of resources can serve as reagents in biomedical experiments, including cell lines, antibodies to detect proteins of interest, and oligonucleotide reagents to analyze specific genomic sequences or transcripts.¹

While it is commonly assumed that antibodies, cell lines, and nucleotide sequence reagents are correctly identified, many studies now indicate that experimental reagents can be wrongly identified in publications. In addition to problems of antibody cross-reactivity and variable performance,² ELISA kits have been found to include antibodies that detect different targets from those claimed.^{3,4} Independent studies have described examples where the identities of some nucleotide sequence reagents do not match their descriptions,⁵⁻¹⁷ including oligonucleotide probes on microarrays,^{18,19} and shRNAs in libraries.²⁰ Finally, due to cross-contamination and/or uncertainty over donor origins, many human cell lines are known to be wrongly identified.²¹⁻²⁵ These wrongly identified reagents can produce a range of downstream consequences, including failed attempts to reproduce published results,^{3,5,7,11,16} and potentially incorrect interpretations of preclinical data leading to misdirected translational research.^{10,12,25}

The capacity to verify reagent identities can help researchers to assess whether published results have been correctly reported. Verifiable reagents include short oligonucleotides and peptides, as their sequences are typically disclosed.^{10,13} Because oligonucleotide and peptide sequences usually cannot be identified by eye, they need to be linked with gene or protein identifiers.^{13,15,17} Reagent identities can therefore be verified by querying nucleotide or amino acid sequences against DNA/RNA or protein sequence databases, and then comparing predicted and claimed reagent identities.^{10,13-15,17} Similarly, cell lines are named using identifiers composed of short

strings of letters and/or numbers.^{22,26} These identifiers can be used to query knowledgebases such as Cellosaurus²⁶ to check whether the claimed identity matches the indexed identity, and whether cell lines are known or suspected to be cross-contaminated or otherwise misclassified. Cell line knowledgebases can also help researchers to navigate ambiguities of cell line nomenclature, such as identifiers that do not reflect the cell line's biological origin, variations in syntax and punctuation, and identical or very similar identifiers being used to denote different cell lines.^{22,26-28} In addition to matching RRIDs to cell line names,²⁸ Cellosaurus also indexes recognized synonyms for cell line identifiers, and in some cases, potentially misspelled identifiers noted in the literature.²⁶

This project sought to extend an earlier study that examined the proportions of human gene research publications that described wrongly identified nucleotide sequences.¹⁵ Park et al.¹⁵ screened >11,700 human gene research papers with the semi-automated tool Seek and Blastn that verifies the identities of nucleotide sequence reagents that are claimed to target human transcripts and genomic sequences.¹³ Papers screened by Seek and Blastn were distributed across five literature corpora, including a corpus of original papers that studied human *miR-145* in preclinical cancer models.¹⁵ Screening 163 *miR-145* papers with Seek and Blastn identified 31 (19%) *miR-145* papers with one or more wrongly identified nucleotide sequences.¹⁵ Other human gene research papers that studied different miRs were also found to include wrongly identified nucleotide sequence reagents.¹⁵ In addition to wrongly identified nucleotide sequences,^{15,17} other issues have been highlighted in the preclinical miR literature, such as experiments where targeted miRs might not be expressed at physiologically relevant levels.²⁹⁻³¹ For example, although *miR-145* has been repeatedly studied in human cancer cell lines of epithelial origin, *miR-145* is not expressed in epithelial cells and would not be expected to serve as a critical regulator in many human cancer cell lines.^{32,33}

To determine whether *miR-145* papers with wrongly identified nucleotide sequences have been published since the analyses of Park et al.,¹⁵ the present study manually verified the identities of nucleotide sequence reagents^{10,17} in *miR-145* papers from 2020 to 2022. We also examined the human cell lines described in recent *miR-145* papers, to determine whether any experiments were conducted with cross-contaminated or misclassified cell lines. These approaches found numerous *miR-145* papers with wrongly identified nucleotide sequences and human cell lines. Furthermore, in two *miR-145* papers with incorrect nucleotide sequences and cell lines, we identified five cell line identifiers that did not correspond to indexed cell lines. Two identifiers (BGC-803 and BSG-823) were recognized misspellings of the contaminated human cell lines, MGC-803 and BGC-823, respectively.²⁶ The remaining identifiers (BSG-823, GSE-1, and TIE-3) were not indexed by Cellosaurus, but could also have represented misspelled versions of contaminated cell lines.²⁶ We collectively referred to these identifiers as being non-verifiable (NV), as their identities were unclear. While all NV identifiers were expected to represent misspellings of similarly named cell lines, the BGC-803, BSG-823, and GSE-1 identifiers appeared to be described as independent cell lines in one *miR-145* paper. As five NV identifiers were found in just two *miR-145* papers, we undertook further analyses to find other NV cell line identifiers and understand how a subset of these identifiers has been described in the literature.

2 | MATERIALS AND METHODS

2.1 | Identification of *miR-145* corpus

To identify recent *miR-145* papers not studied by Park et al.,¹⁵ Web of Science was searched using the following parameters: Title = “*miR-145*” AND Topic = “Cancer” AND All Fields = “Human” AND Document Type = “Article” for articles published from January 1, 2020 to September 1, 2022. Search results were exported into Google Sheets, articles were visually screened to ensure that each paper had studied human *miR-145*, and the title, journal, publisher, publication year, PubMed ID (PMID), and journal impact factor (JIF)³⁴ for the publication year were recorded. Human cancer type(s) studied were identified by visual inspection of text. Country of origin was determined according to the affiliations of at least half of the authors.^{15,17} Publications were judged to be hospital-affiliated^{15,17} or university-affiliated if at least half of the authors listed relevant affiliations. In the case of no majority, the first author's affiliation(s) determined affiliation status.¹⁷

2.2 | Verification of nucleotide sequence reagent identities

Nucleotide sequence identities were verified as described in the Supporting Information S1.^{13-15,17,35,36}

2.3 | Verification of cell line identities

Cell line identifiers in *miR-145* papers were extracted using copy/paste functions and used to query Cellosaurus.²⁶ Cell lines indexed by Cellosaurus as being (potentially) contaminated with a different cell line or otherwise misclassified were flagged if the article did not recognize (i) the cell line's contaminated/misclassified status or (ii) the use of a contaminated/misclassified cell line as a study limitation. Where Cellosaurus did not index an identifier as a human cell line, or only recognized an identifier as a misspelled human cell line identifier,²⁶ the identifier was referred to as NV. The first papers that our team identified that described individual NV identifier(s) were referred to as index papers.

2.4 | Literature searches employing single cell line identifiers

Cell line identifiers were employed as keywords to search Google Scholar (Figures S1 and S2). Identifiers were written with and without a dash (“-”) between alphabetic and numeric identifier components, where “-” was also recognized as a space by Google Scholar. Where search results and GeneCards³⁵ indicated that any cell line identifier corresponded to a human gene,^{27,37} identifiers were combined with the terms “cells” and “cell line,” that is, “TIE-3 cells,” “TIE-3 cell line.”

Publications retrieved by NV cell line identifiers (± “cells” and “cell line”) were prioritized for analysis according to their best-match ranking. Where NV cell line identifiers retrieved up to 50 sources, all publications were selected for further analysis (Figure S2). Otherwise, 50–200 publications were selected from the first 6–20 pages of Google Scholar results, with a focus on top-ranked results (Figure S2).

Publications were further analyzed if they could be downloaded in full or sourced through University of Sydney library services and were written in English. Text search functions were used to identify the queried cell line identifier, and all figures and tables were visually inspected. If publications included the queried identifier, the following information was extracted: (i) article type, (ii) PMID, DOI or other identifier, (iii) publication year, (iv) journal, (v) publisher, (vi) JIF for the publication year (if available), (vii) publication title, (viii) country affiliation, (ix) hospital/ university affiliation, (x) details of any post-publication notices (corrections, expression of concerns, retractions), (xi) whether the queried identifier was used in experiments and/or referenced, (xii) the claimed cell type (e.g., gastric cancer), (xiii) whether the queried identifier was listed with other cell lines, and their identities, (xiv) cell line sources, if provided, and (xv) whether cell line identities were checked using short tandem repeat (STR) profiling.

Google Scholar results were also triaged by date, to identify the earliest publications that had referred to queried identifiers (Figures S1 and S2). Publications were subjected to further analyses if they could be sourced as described above, and if at least the abstract was written in English and mentioned the queried cell line identifier. Publications were visually inspected for descriptions of cell line establishment^{1,23} and other publication features were recorded as described above.

2.5 | Combined literature analyses of NV identifiers and similarly named human cell lines

Google Scholar searches were conducted with NV identifiers paired with known human cell line identifiers, where the NV identifier was either a recognized misspelling of a human cell line identifier²⁶ and/or was similar to and/or occurred in association with a similarly named human cell line (Figures S1 and S2). One index paper listed two NV identifiers (BGC-803 and BSG-823) with two similarly named cell lines (MGC-803 and BGC-823). This identifier list (HS-746T, BSG-823, MKN-28, 9811, BGC-803, MGC-803, and BGC-823) was also used as a search term (Data file S1). Publications were triaged and analyzed as described above. Where cell line sources represented repositories with accessible online catalogs, catalogs were queried with NV cell line identifiers written (i) as single words and (ii) with a space or “-” between the alphabetic and numeric components. For each queried catalog, a human cell line identifier (e.g., MCF-7) was employed as a positive control.

As some publications that referred to NV cell line identifiers stated that cell line identities had been confirmed using STR profiling, additional Google Scholar searches combined NV cell line identifiers with the terms “STR” and “short tandem repeat”. All publications that referred to NV cell line identifiers were searched for references to STR profiling and STR profiling results.

As some *miR-145* papers referred to NV cell line identifiers as independent cell lines (see Results), publications that referred to NV and similarly named human cell line identifiers were inspected to determine whether NV cell line identifiers were (i) likely to represent misspelled versions of human cell line identifiers, or (ii) referred to as independent cell lines. An NV cell line identifier was indicated to represent a *misspelling* if the NV identifier (i) was used alternately with similarly named human cell line(s) such that identifiers were never directly connected, either in the same sentence (e.g., “and”) or adjacent sentences (e.g., “also”, “in contrast”); (ii) did not appear in any list of cell lines with any similarly named cell line; and (iii) were not included in any single experiment with any similarly named cell line. In contrast, NV cell line identifiers were indicated to represent *independent cell lines* if (i) the NV identifier was used in the absence of any similarly-named human cell line(s); (ii) NV and similarly-named human cell line identifiers were included in any list of cell lines studied in the paper; (iii) results for NV and similarly-named human cell line(s) were shown in the same experiment(s); and/or (iv) NV and similarly-named human cell line identifiers were directly connected, either in the same or adjacent sentences.

2.6 | Representation of cell line identifiers

We have chosen to show all cell line identifiers in the text with a dash (-) between alphabetical and numeric components, and to list identifiers in alphabetical order where possible, for improved readability.

2.7 | Citation analyses

Google Scholar citations were collected on January 29, 2024.

3 | RESULTS

3.1 | miR-145 corpus and nucleotide sequence analyses

Querying Web of Science retrieved 36 original papers published between January 1, 2020 and September 1, 2022 that were visually confirmed to include human *miR-145* in their titles and refer to human cancer. Most (24/36, 67%) *miR-145* papers described nucleotide sequence reagents, where these 24 *miR-145* articles were published in 21 journals and examined *miR-145* in the context of 15 human cancer types (Table S1).

The 24 *miR-145* papers described 339 nucleotide sequence reagents, of which almost one quarter (83/339, 24%) were predicted to be wrongly identified (Table 1). Most incorrect nucleotide sequences were claimed targeting reagents that were either predicted to be non-targeting in humans (43/83, 52%), or to target a different human gene or genomic sequence from that claimed (39/83, 47%) (Table S2, Data file S2). Wrongly identified nucleotide sequences were found in over half (20/36, 56%) of *miR-145* papers, and in most (20/24, 83%) *miR-145* papers that described nucleotide sequence reagents, with a median of 3 (range 1–16) wrongly identified sequences/paper (Table 1). The 20 *miR-145* papers with wrongly identified nucleotide sequence(s) had been cited 382 times by January 2024 (Table S1).

3.2 | Cross-contaminated and misclassified cell lines

A total of 21 *miR-145* papers described experiments in cell lines, all of which also described nucleotide sequence(s) (Table S1). These 21 papers referred to 91 different cell lines, with a median of 5 (range 1–13) cell lines/paper (Table 1), where some cell lines were described in multiple *miR-145* papers (Data file S2). Fourteen percent (15/107) of all cell line descriptions represented wrongly identified cell lines (Table 1) and almost one third (6/21, 29%) of *miR-145* papers that described cell lines included at least one wrongly identified cell line. All wrongly identified cell lines were claimed human cancer cell lines, where most have been found to be contaminated by HeLa and/or other cancer cell lines^{38–41} (Table S3).

3.3 | Identification of NV cell line identifiers

Two *miR-145* papers that described wrongly identified nucleotide sequence(s) and cell line(s) also described five cell line identifier(s) that were either recognized misspellings of contaminated cell lines,²⁶

TABLE 1 Verification of reagent identities in human *miR-145* papers published from 2020 to 2022.

	Nucleotide sequence reagents	Cell lines ^a	Total
Number of <i>miR-145</i> papers with reagents whose identities could be verified	<i>n</i> = 24	<i>n</i> = 21	<i>n</i> = 24
Number of reagents per <i>miR-145</i> paper, median (range)	13 (2–54)	5 (1–13)	18 (4–56)
Number of <i>miR-145</i> papers according to country of origin	China (<i>n</i> = 21), Brazil (<i>n</i> = 1), Taiwan (<i>n</i> = 1), Thailand (<i>n</i> = 1)	China (<i>n</i> = 19), Taiwan (<i>n</i> = 1), Thailand (<i>n</i> = 1)	China (<i>n</i> = 21), Brazil (<i>n</i> = 1), Taiwan (<i>n</i> = 1), Thailand (<i>n</i> = 1)
Proportion (percentage) of wrongly identified reagents	83/339 (24%)	15/107 (14%)	98/446 (22%)
Proportion (percentage) of <i>miR-145</i> papers with wrongly identified reagents	20/24 (83%)	6/21 (29%)	20/24 (83%)
Number of wrongly identified reagents per paper, median (range)	3 (1–16)	3 (1–4)	5 (1–16)
Number of papers with wrongly identified reagents according to country of origin	China (<i>n</i> = 19), Thailand (<i>n</i> = 1)	China (<i>n</i> = 6)	China (<i>n</i> = 19), Thailand (<i>n</i> = 1)
Proportion (percentage) of papers with wrongly identified reagents that were affiliated with hospitals	16/20 (80%)	5/6 (83%)	16/20 (80%)

^aRefers to cell lines indexed by Cellosaurus (release 47, October 2023), does not include non-verifiable cell line identifiers.

and/or not indexed by Cellosaurus (Tables S4 and S5). Two NV identifiers (BGC-803 and BSG-823) were recognized misspellings²⁶ of the contaminated cell lines BGC-823 or MGC-803,^{38–41} whereas the BSG-803, GSE-1, and TIE-3 identifiers were similar to the contaminated cell lines BGC-823/MGC-803,^{38–41} GES-1,^{39,41} and TE-3, respectively.⁴² We commonly referred to recognized misspelled and non-indexed identifiers²⁶ as NV, as their identities appeared uncertain.

As five NV identifiers were found in two *miR-145* papers, we reasoned that other NV cell line identifiers might exist in the literature. Google Scholar searches successively identified 23 different NV cell line identifiers (Tables S4 and S5), many of which were similar to contaminated cell line identifiers such as BGC-823, MGC-803, or SGC-7901 (Table S5). The frequencies of these NV identifiers varied. Some were found only once as cell line identifiers, whereas others were found in hundreds of Google Scholar sources (Table S5).

We analyzed eight NV identifiers in detail (BGC-803, BSG-803, BSG-823, GSE-1, HGC-7901, HGC-803, MGC-823, and TIE-3), including the five NV identifiers found in *miR-145* papers (Tables S4 and S5). Four NV identifiers (BGC-803, BSG-823, GSE-1, and HGC-7901) were referred to as independent cell lines in the corresponding index papers, whereas the remaining four NV identifiers appeared to represent misspellings of known cell lines (Table S4). Furthermore, three NV identifiers (BGC-803, BSG-823, and GSE-1) were claimed to have been obtained from a cell line repository with an online catalog (Table S4), yet querying this repository catalog did not identify these three cell line identifiers.

3.4 | Descriptions of NV identifiers in the research literature

We assessed 420 articles that referred to at least one of the eight NV identifiers (Data file S1). This revealed that in 185/420 (44%) papers, all eight NV identifiers could represent misspellings of similarly named cell lines, which were typically HeLa-contaminated cancer cell lines.^{38–41} For example, in the 79 papers that referred to both GSE-1 and GES-1, GSE-1 appeared to consistently represent a misspelling of GES-1 (Data file S1). Similarly, in the single paper that referred to TIE-3 as a cell line identifier, TIE-3 appeared to be a misspelling of TE-3 (Table S4).

While many instances of NV identifiers represented misspellings, more than half (235/420, 56%) of papers appeared to describe at least one of seven NV identifiers (BGC-803, BSG-803, BSG-823, GSE-1, HGC-7901, HGC-803, and/or MGC-823) as independent cell lines (Tables 2 and 3, Data file S1). Some original papers that described the results of experiments involving NV cell line identifier(s) referred to these identifier(s) without reference to any similarly named human cell line.²⁶ For example, 34 original papers referred to GSE-1 cells without mentioning GES-1 (Table 2). In other original papers, NV and similarly named cell lines were (i) included in lists of cell lines employed in experiments (e.g., “The human gastric cancer cell lines ... MGC-823, MGC-803...were used in this study”), (ii) used in the same experiments according to information provided in the text, and/or (iii) shown together as results for the same experiment(s) in figures and/or tables (Data file S1). The numbers of original papers that described NV identifiers as independent cell lines ranged from one paper (BSG-803) to

TABLE 2 Original papers that describe experiments where non-verifiable (NV) cell line identifier(s) were described as independent cell line(s).

NV cell line identifier	Original papers referring to NV identifier as independent cell line ^a	Publication years, range	Number of individual journals/publishers	Journal Impact Factor, range	Most frequent country of origin proportion (%)	Most frequent institution type proportion (%)	Cell line repository sources	Papers referring to derivation of STR profiles, proportion (%)
BGC-803	n = 116	2006–2023	n = 82/n = 34	0.2–12.7	China 115/116 (99%)	Hospital 73/116 (63%)	Named collaborator/institute/laboratory; ATCC; BeNa Culture Collection (Beijing, China); Cell Bank of the Chinese Academy of Sciences; Type Culture Collection of Chinese Academy of Sciences	4/116 (3%)
BSG-803	n = 1	2020	n = 1/n = 1	4.1	China 1/1 (100%)	Hospital 1/1 (100%)	Not stated	0/1 (0%)
BSG-823	n = 14	2015–2022	n = 13/n = 11	2.5–6.4	China 14/14 (100%)	Hospital 12/14 (86%)	Named collaborating institute; Cell Bank of the Chinese Academy of Sciences; National Infrastructure of Cell Line Resources of China	0/14 (0%)
GSE-1	n = 34	2004–2023	n = 30/n = 16	2.0–9.7	China 33/34 (97%)	Hospital 27/34 (79%)	Named collaborating institute; ATCC; BeNa Culture Collection; China Center for Type Culture Collection	2/34 (6%)
HGC-803	n = 3	2016–2018	n = 3/n = 3	1.7–4.3	China 3/3 (100%)	Hospital 3/3 (100%)	ATCC; Cell Bank of the Chinese Academy of Sciences	0/3 (0%)
HGC-7901	n = 7	2015–2023	n = 7/n = 5	2.1–8.5	China 7/7 (100%)	Hospital 4/7 (57%)	Cell Bank of Chinese Academy of Sciences	0/7 (0%)
MGC-823	n = 34	2005–2023	n = 30/n = 16	0.3–11.2	China 33/34 (97%)	Hospital 23/34 (68%)	Named collaborator or collaborating institute; ATCC; Cell Bank of Chinese Academy of Sciences; China Center for Type Culture Collection	5/34 (15%)

^aOriginal papers that referred to experiments that employed NV cell line(s) (i) without reference to any similarly-named cell line and/or (ii) with a similarly-named cell line, such that NV cell line(s) were referred to as independent cell line(s).

TABLE 3 Literature reviews, commentaries, book chapters, original articles, and preprints where non-verifiable (NV) cell line identifier(s) were referred to independent cell line(s) in the absence of experimental results.

NV cell line identifier	Number of papers (literature reviews) referring to NV identifier as independent cell line	Publication years, range	Number of individual journals/publishers	Journal Impact Factor, range	Most frequent country of origin proportion (%)	Most frequent institution type proportion (%)
BGC-803	$n = 27$ ($n = 22$)	2009–2023	$n = 22/n = 8$	1.8–13.6	China 11/21 (41%)	University 22/27 (81%)
BSG-823	$n = 8$ ($n = 7$)	2016–2023	$n = 7/n = 5$	2.8–13.0	China 4/8 (50%)	University 5/8 (63%)
GSE-1	$n = 2$ ($n = 2$)	2023	$n = 2/n = 2$	No information	N/A ^a	University 2/2 (100%)
HGC-7901	$n = 1$ ($n = 1$)	2020	$n = 1/n = 1$	6.2	China 1/1 (100%)	Hospital 1/1 (100%)
HGC-803	$n = 4$ ($n = 2$)	2019–2023	$n = 3/n = 3$	1.1–6.6	China 2/4 (50%)	N/A ^a
MGC-823	$n = 2$ ($n = 2$)	2019–2021	$n = 2/n = 2$	7.8–16.8	N/A ^a	University 2/2 (100%)

^aN/A in relation to country/institution indicates no majority.

116 papers (BGC-803), where not all BGC-803 papers were analyzed. Papers that did not describe experiments with NV cell lines, including 36 literature reviews, also referred to NV identifiers as independent cell lines (Table 3). In these papers, NV identifiers were found in the absence of any similarly named cell line or were paired with similar cell line identifiers, either in literature review text, or in introduction or discussion sections of the original papers (Data file S1).

Papers describing NV identifiers as independent cell lines were published between 2004 and 2023 in fields including cancer research, biochemistry, chemistry, and drug discovery, and in journals with impact factors ranging from 0.2 to 16.8 (Tables 2 and 3). Although five NV identifiers were found in *miR-145* papers, only 25% (58/235) of papers that described NV cell lines referred to miRs in their titles. Almost all original papers that described experiments involving NV cell lines were published by authors in China, where most publications were affiliated with hospitals (Table 2). In contrast, papers such as literature reviews that referred to NV cell lines in the absence of experimental results were mostly affiliated with universities (Table 3).

3.5 | NV cell line establishment, sources, and STR profiling results

To clarify the origins of NV cell line identifiers, we analyzed the earliest publications that mentioned NV identifiers and/or referred to NV identifiers as cell lines. The earliest papers that mentioned NV cell line identifiers were published between 1988 and 2020, whereas the first papers that referred to NV identifiers as independent cell lines were published between 2004 and 2020 (Figure 1). The BSG-803, BSG-823, GSE-1, and HGC-7901 identifiers appeared to be first mentioned as independent cell lines, whereas BGC-803, HGC-803, and MGC-823 were referred to as independent cell lines 1–17 years after they were first mentioned in the literature (Figure 1). No earliest paper that we could find described how the seven NV cell lines were established or referenced cell line establishment papers.

As we did not find three NV identifiers in the described cell line repository in one index paper (Table S4), we extracted all cell line sources described in the 420 papers analyzed (Data file S1). In

papers where NV identifiers were referred to as independent cell lines, four NV cell lines (BGC-803, GSE-1, HGC-803, and MGC-823) were described as sourced from the American Type Culture Collection (ATCC).⁴³ Six NV cell lines (BGC-803, BSG-823, GSE-1, HGC-7901, HGC-803, and MGC-823) were described as sourced from the Chinese National Infrastructure of Cell Line Resource,⁴⁴ and three NV cell lines (BGC-803, GSE-1, and MGC-823) were described as sourced from the China Center for Type Culture Collection.^{38–41} Searching the online catalogs of these three repositories did not identify the claimed NV identifiers or cell lines (Table S5).

A minority (16/420, 4%) of papers that mentioned NV cell line identifiers described the use of STR profiling to confirm cell line identities (Table 2, Data file S1). These 16 papers either described STR profiling to confirm the identities of named NV cell lines or included overarching statements that were presumed to refer to all cell lines. Although 11/16 papers referred to three NV cell line identifiers (BGC-803, GSE-1, and/or MGC-823) as independent cell lines, none of these papers either showed or described STR profiling results for any NV cell line (Data file S1).

4 | DISCUSSION

This study commenced by checking nucleotide sequence reagent identities in a cohort of human *miR-145* papers. Having previously identified wrongly identified nucleotide sequences in 19% of human *miR-145* papers using Seek and Blastn screening,¹⁵ manual screening revealed at least one wrongly identified nucleotide sequence in over half (56%) *miR-145* papers published between 2020 and 2022, and in 83% *miR-145* papers that described nucleotide sequences (Table 1). Higher proportions of papers with wrongly identified nucleotide sequences likely reflect the use of manual screening,¹⁷ indicating that semi-automated screening could underestimate the numbers of gene research papers with wrongly identified nucleotide sequences.¹⁵

Just under one third of relevant *miR-145* papers also described wrongly identified cell line(s) (Table 1). Unexpectedly, two *miR-145* papers that described wrongly identified nucleotide sequence(s) and cell line(s) also included NV cell line identifiers, where some were

Timeline of non-verifiable identifier use

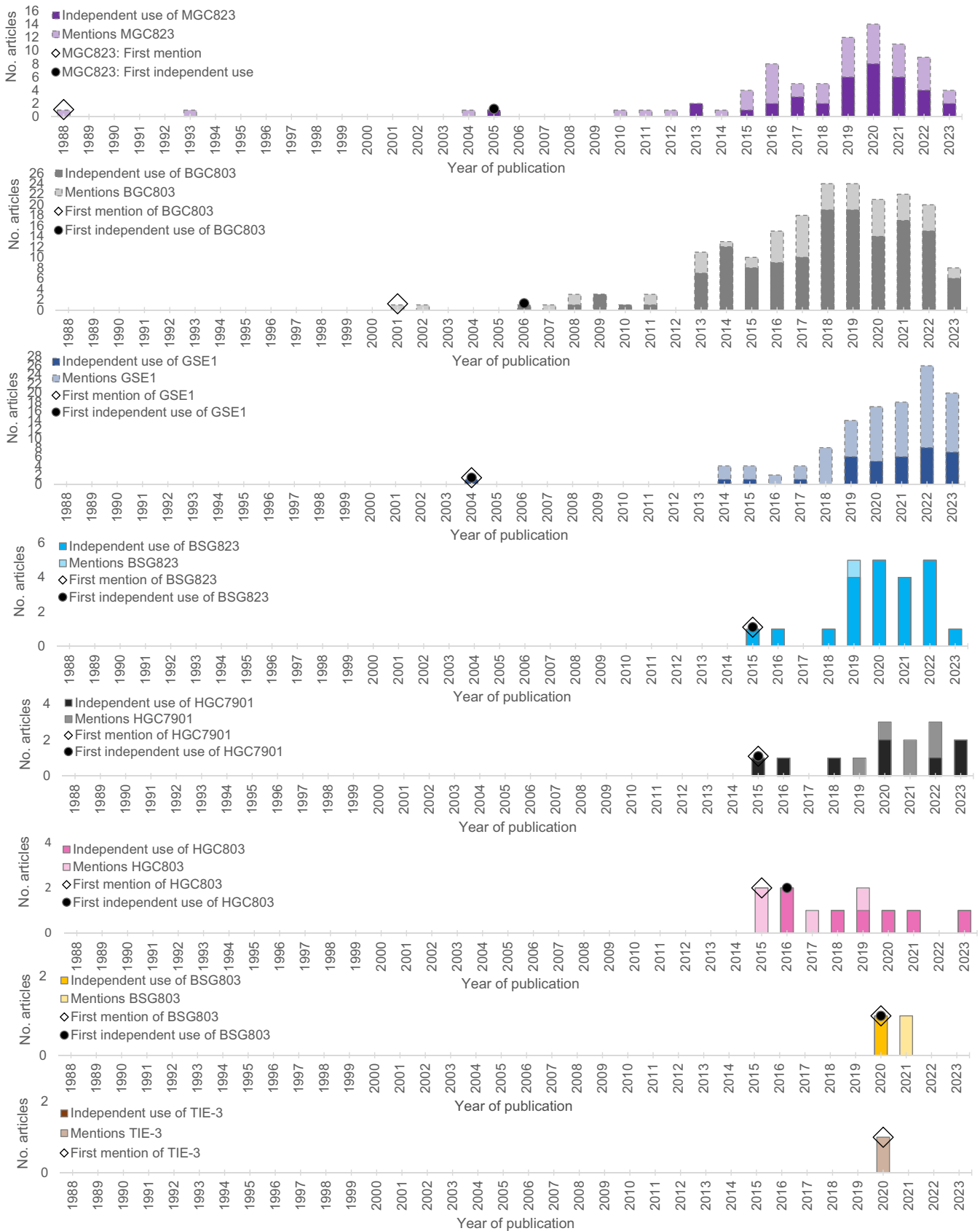


FIGURE 1 Legend on next page.

referred to as independent cell lines. As five NV cell line identifiers were found in two *miR-145* papers, and paralleled previous descriptions of NV circular RNAs,^{17,45} we sought to understand whether other NV cell line identifiers could be found in the literature and how a subset of these NV identifiers have been described. Studying eight NV cell line identifiers across 420 papers revealed that all eight NV identifiers could represent identifier misspellings in at least some publication(s). However, 235 (56%) papers unexpectedly described seven NV identifier(s) (BGC-803, BSG-803, BSG-823, GSE-1, HGC-7901, HGC-803, and MGC-823) as if they were independent cell lines. These NV identifiers have been mentioned in the literature as early as 1988 and have been referred to as independent cell lines since 2004 (Figure 1) in fields ranging from drug discovery and testing to gene research. Nonetheless, we could not find any papers that described how these seven NV cell lines were established. Six NV cell lines were stated to have been sourced from externally accessible cell line repositories, including ATCC, yet checking their online catalogs did not identify these NV cell lines. We also could not identify any published STR profiles for BGC-803, GSE-1, and MGC-823 cells, despite papers stating that these STR profiles had been confirmed. These NV cell lines were also not described by studies describing STR profiles for similarly named contaminated cell lines.^{38–41}

4.1 | Limitations

We recognize that this study has several limitations. Although we attempted to study all papers that mentioned the identifiers BSG-803, BSG-823, HGC-803, HGC-7901, and TIE-3, we may have still missed some relevant publications. For example, as GSE-1 and TIE-3 also represent human genes, our use of additional search terms to restrict results may have excluded relevant papers. We conducted literature searches for NV cell line identifiers during 2023, and so we may have missed some relevant papers published late in 2023 (Figure 1). We also recognize that as we only screened papers with at least some English text, we could have failed to identify papers that described how NV cell lines were established, and/or the results of STR profiling. Due to numbers of available publications, we did not analyze all papers that mentioned the BGC-803, GSE-1, and MGC-823 identifiers. We did not study all NV identifiers that we found (Table S5) and we made no attempt to find all NV cell line identifiers that might exist in the literature.

While claimed NV cell lines could not be found in three external repositories with online catalogs, including ATCC,⁴³ we recognize that stocks of these and other NV cell lines might be held elsewhere, including repositories lacking online catalogs. We also recognize that

at least some papers that referred to NV identifiers without reference to similarly named cell line(s) could have employed the NV identifier as a misspelling, and that similar explanations could apply to some other descriptions of independent NV cell lines. Nonetheless, given that many NV cell line identifiers could represent misspellings of cross-contaminated cancer cell lines (Table S5), descriptions of any intended cell lines could remain problematic.

4.2 | Significance and next steps

Our results indicate that NV cell line identifiers can gain new identities as independent cell lines without published descriptions of how cell lines were established, STR profiles, or making cell lines available through external suppliers, a process that we refer to as “miscelling” (Figure 2). Whereas some NV cell lines appear to arise de novo, others are predated by cell line identifier misspellings (Figure 1), suggesting that misspelled identifiers may predispose to subsequent NV cell line descriptions (Figure 2). Identifier misspellings and NV cell lines could lead some researchers to misspell cell line identifiers when carrying out their research. Descriptions of NV identifiers in *miR-145* publications with wrongly identified nucleotide sequence(s) could also indicate experiments that were not performed as described.¹⁵ For example, NV cell lines could also arise through research paper mills confusing misspelled cell line identifiers for independent cell lines, due to lack of time, expertise, and attention to detail (Figure 2).

Our results also show that NV cell lines are published across different research fields, reflecting the widespread research use of human cancer cell lines. However, where cell line origins and identities are unknown, any resulting data cannot be interpreted or translated. If NV cell lines cannot be sourced from external repositories, their claimed identities cannot be verified, and published research cannot be reproduced. Despite anticipated challenges in sourcing NV cell lines, some researchers might still attempt to reproduce results using other cell lines, leading to wasted time and resources. We therefore recommend that NV cell line identities be clarified as soon as possible, by disclosing existing STR profiles and supplying cell line stocks to independent groups for STR profiling and phenotypic testing. While we could not identify sources for NV cell lines, teams that have described these cell lines could provide samples for testing, where dates of cell line stocks should predate published experiments. Testing cell line stocks from different sources would have the added advantage of allowing STR profiles for multiple cell line stocks to be directly compared.^{38–41}

Repeated references to misspelled cell line identifiers in Cellosaurus,²⁶ combined with the many NV identifiers found in this

FIGURE 1 Bar graphs showing numbers of papers (Y-axis) that refer to the indicated NV cell line identifiers (shown at top left) per year (X-axis). Light-shaded bars indicate papers that mention NV cell line identifiers as likely misspellings, whereas darker bars indicate papers in which NV cell line identifiers were referred to as independent cell lines. Broken lines indicate publication numbers derived from studying a subset of papers (BGC-803, GSE-1, and MGC-823), as described in the Methods. Open diamonds indicate the earliest publication that mentioned each identifier, whereas filled circles indicate the earliest paper to refer to an NV identifier as an independent cell line. Graphs are ranked from the top according to years of earliest publication (1988–2020). NV, non-verifiable.

study, suggest that other NV cell line identifiers remain to be discovered. Misspelled and other NV cell line identifiers may be difficult to notice through reading, particularly where cell line identifiers are non-intuitive and can be written with different punctuation and/or syntax^{22,26–28} (Figure 2). Furthermore, as misspelled versions of familiar commercial logos can be recognized as if they were correct,⁴⁶ even expert readers might not notice subtle identifier misspellings. Most NV identifiers described by this study were found by undergraduate students working on short-term publication integrity

projects. Similar projects could be conducted in the future, particularly as publication integrity projects can be scaled according to available time and resources and are suitable as individual or group projects.

We suggest steps to reduce descriptions of NV cell lines (Table 4), where some have also been proposed to reduce descriptions of contaminated or misidentified cell lines.^{23,26,28,47–49} To raise broad awareness of NV cell lines, we suggest creating a dynamic, openly accessible registry of NV and misspelled cell lines. A comprehensive list of NV cell lines and misspelled cell line identifiers would allow these to be

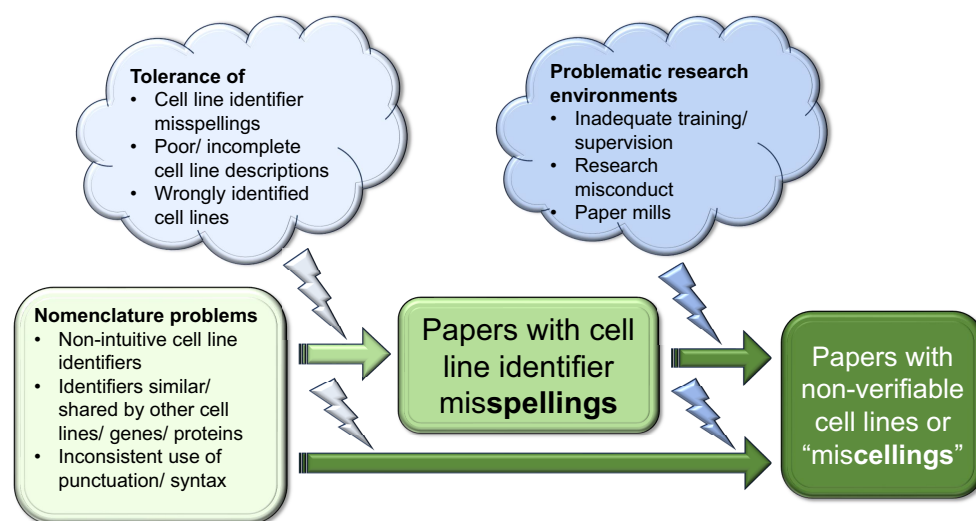


FIGURE 2 Summary of factors that could predispose to cell line identifier misspellings and NV cell lines or “miscellings,” recognizing that papers referring to NV cell lines can either appear de novo, or follow publications with NV cell line identifier misspellings. NV, non-verifiable.

TABLE 4 Actions to prevent descriptions of non-verifiable (NV) cell lines in research publications.

Actions to reduce descriptions of NV cell lines	Previously described in context of wrongly identified cell lines ^a	Stakeholders
Establish dynamic register of misspelled cell line identifiers and NV cell lines		Researchers, research funding organizations
Create screening tools/plugin-ins to detect misspelled, NV, and wrongly identified cell lines	28, 50	
Insist on use of correct cell line identifiers in research publications. Zero tolerance for misspelled cell line identifiers.	22, 47, 48	Authors, journals, publishers, peer reviewers
Cell lines to be described in easily searchable publication sections, for example, title and/or abstract	23	
Cell lines to be described by at least two identifiers, for example, cell line identifier + RRID	24, 28, 48	
All source(s) of published cell lines to be fully disclosed	23	
All cell lines described in results to also be clearly described in the methods		
Publications describing new cell line to include description of donor origin, method of establishment, culture conditions, phenotyping, and genotyping data including STR profile	1, 23	
Verification of all cell line identities prior to peer review	23, 24	Journals, publishers
Immediate publication of expressions of concern or editorial notes for papers that describe or refer to NV cell lines	23, 49	
Retraction of original papers describing experiments with cell lines whose identities cannot be verified		
Published corrections to papers such as literature reviews that refer to cell lines whose identities cannot be verified		

^aReferences that have described or proposed similar approach(es) to reduce descriptions of cross-contaminated or misidentified cell lines.

included in screening tools used by publishers and researchers, where the seven NV cell lines described in the present study have been added to the Problematic Paper Screener.⁵⁰ As we identified multiple NV cell line identifiers in some papers, research application of literature screening tools should also accelerate the discovery of other NV identifiers.

Descriptions of NV cell lines could also be reduced by improving the quality of cell line reporting (Table 4). For example, requiring authors to include the corresponding RRID²⁸ for all cell lines could reduce descriptions of NV cell lines. At the same time, as journal guideline implementation is typically not associated with clear improvements in reagent reporting,⁵¹ other actions will be required. Given the uncertainty surrounding the origins and status of NV cell lines, journals and publishers need to show a zero-tolerance approach towards manuscripts and publications that describe misspelled and NV cell lines. Screening for NV cell lines in submitted manuscripts should be included in publisher workflows, where flagged manuscripts should not be sent for peer review. It has been recommended that publications that describe experiments with misidentified cell lines should be flagged with editorial notes or expressions of concern.^{23,49} This should extend to all publications that have described NV cell lines. If the identities and origins of cell line(s) cannot be verified, publications that have described experiments with NV cell lines should be considered for retraction, and other papers referring to NV cell lines, such as literature reviews, should be flagged through corrections (Table 4).

4.3 | Summary

As reported for other identifiers that lack intrinsic visual sense,⁵² cell line identifiers are prone to being wrongly written or misspelled.^{22,26} Although some NV identifiers are likely to represent misspellings of similarly named cell lines, our results indicate that some misspelled cell line identifiers can gain new identities as independent cell lines, a process that we describe as “miscelling.” As NV cell lines represent challenges to research integrity and reproducibility, further research is needed to clarify the status and identities of NV cell lines and whether other NV cell lines have been described. In the meantime, publications describing NV cell lines call for greater focus on improving published descriptions of human cell lines, to ensure the transparency of cell line models employed in biomedical research.

AUTHOR CONTRIBUTIONS

Conceptualization: JAB, GC, and CL. Methodology: JAB, DJO, PP, ACD, GC, and CL. Formal analysis: DJO, PP, RAKR, GJ, YA, MDA, NRE, MH, JK, AL, and JAB. Writing—original draft preparation: JAB and DJO. Writing—review and editing: DJO, PP, RAKR, GJ, YA, MDA, NRE, MH, JK, AL, JD, GC, CL, ACD, and JAB. Funding acquisition: JAB, ACD, and CL. Supervision: JAB, JJD, and TS. The work reported in the paper has been performed by the authors, unless clearly specified in the text.

ACKNOWLEDGMENTS

JAB, ACD, and CL gratefully acknowledge funding from the National Health and Medical Research Council of Australia (NHMRC) Ideas Grant

ID APP1184263. PP is supported by a Research Training Program scholarship at the University of Sydney. RAKR gratefully acknowledges support from the Dr John N. Nicholson fellowship from Northwestern University and Moderna Inc. “Identifying bias and improving reproducibility in RNA-seq computational pipelines.” JJD is supported by an NHMRC Investigator Grant (GNT2008066). TS is supported by NIH (USA) Grant AG068544. We are grateful to two anonymous reviewers for their insightful comments and suggestions that have improved our manuscript. We apologize to authors whose publications were not cited due to limits on the number of references. Open access publishing facilitated by The University of Sydney, as part of the Wiley - The University of Sydney agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data generated or analyzed during this study are included in the manuscript and its Supplementary Information files. All information extracted from or about analyzed publications, as well as Google Scholar citation data, is available within the public domain. Further information is available from the corresponding author upon request.

ORCID

Jennifer A. Byrne  <https://orcid.org/0000-0002-8923-0587>

REFERENCES

- Vasilevsky NA, Brush MH, Paddock H, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. *PeerJ*. 2013;1:e148.
- Ayoubi R, Ryan J, Biddle MS, et al. Scaling of an antibody validation procedure enables quantification of antibody performance in major research applications. *eLife*. 2023;12:RP91645.
- Rodland KD. As if biomarker discovery isn't hard enough: the consequences of poorly characterized reagents. *Clin Chem*. 2014;60:290-291.
- Prassas I, Brinc D, Farkona S, et al. False biomarker discovery due to reactivity of a commercial ELISA for CUZD1 with cancer antigen CA125. *Clin Chem*. 2014;60:381-388.
- Habbal W, Monem F, Gärtner BC. Errors in published sequences of human cytomegalovirus primers and probes: do we need more quality control? *J Clin Microbiol*. 2005;43:5408-5409.
- Katavetin P, Nangaku M, Fujita T. Wrong primer for rat angiotensinogen mRNA. *Am J Physiol Renal Physiol*. 2005;288:F1078.
- Chiarella P, Carbonari D, Iavicoli S. Utility of checklist to describe experimental methods for investigating molecular biomarkers. *Biomark Med*. 2015;9:989-995.
- Kocemba KA, Dudzik P, Ostrowska B, Laidler P. Incorrect analysis of MCAM gene promoter methylation in prostate cancer. *Prostate*. 2016;76:1464-1465.
- Tamm A. Incorrect primer sequences in the article on methylprednisolone treatment. *Acta Neurol Scand*. 2016;134:90.
- Byrne JA, Labbé C. Striking similarities between publications from China describing single gene knockdown experiments in human cancer cell lines. *Scientometrics*. 2017;110:1471-1493.
- Khoshi A, Sirghani A, Ghazisaeedi M, Mahmudabadi AZ, Azimian A. Association between TPO Asn698Thr and Thr725Pro gene polymorphisms and serum anti-TPO levels in Iranian patients with subclinical hypothyroidism. *Hormones*. 2017;16:75-83.

12. Bustin S, Nolan T. Talking the talk, but not walking the walk: RT-qPCR as a paradigm for the lack of reproducibility in molecular research. *Eur J Clin Invest*. 2017;47:756-774.
13. Labbé C, Grima N, Gautier T, Favier B, Byrne JA. Semi-automated fact-checking of nucleotide sequence reagents in biomedical research publications: the Seek & Blastn tool. *PLoS One*. 2019;14:e0213266.
14. Byrne JA, Park Y, West RA, Capes-Davis A, Cabanac G, Labbé C. The thin ret(raction) line: biomedical journal responses to reports of incorrect non-targeting nucleotide sequence reagents in human gene knockdown publications. *Scientometrics*. 2021;126:3513-3534.
15. Park Y, West RA, Pathmendra P, et al. Identification of human gene research articles with wrongly identified nucleotide sequences. *Life Sci Alliance*. 2022;5:e202101203.
16. Zielske SP, Cackowski FC. Critical analysis of the hypothesized SNHG1/miR-195-5p/YAP1 axis. *Funct Integr Genomics*. 2023;23:2.
17. Pathmendra P, Park Y, Enguita FJ, Byrne JA. Verification of nucleotide sequence reagent identities in original publications in high impact factor cancer research journals. *Naunyn Schmiedebergs Arch Pharmacol*. 2024; online ahead of print.
18. Harbig J, Sprinkle R, Enkemann SA. A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucleic Acids Res*. 2005;33:e31.
19. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*. 2006;22:101-109.
20. Perkins LA, Holderbaum L, Tao R, et al. The transgenic RNAi project at Harvard Medical School: resources and validation. *Genetics*. 2015; 201:843-852.
21. Capes-Davis A, Theodosopoulos G, Atkin I, et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer*. 2010;127:1-8.
22. Freedman LP, Gibson MC, Ethier SP, Soule HR, Neve RM, Reid YA. Reproducibility: changing the policies and culture of cell line authentication. *Nat Methods*. 2015;12:493-497.
23. Horbach SP, Halfman W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. *PLoS One*. 2017;12: e0186281.
24. Souren NY, Fusenig NE, Heck S, et al. Cell line authentication: a necessity for reproducible biomedical research. *EMBO J*. 2022;41:e111307.
25. Korch CT, Capes-Davis A. The extensive and expensive impacts of HEP-2 [HeLa], intestine 407 [HeLa], and other false cell lines in journal publications. *SLAS Discov*. 2021;26:1268-1279.
26. Bairoch A. The Cellosaurus, a cell-line knowledge resource. *J Biomol Tech*. 2018;29:25-38.
27. Kafkas Ş, Sarntivijai S, Hoehndorf R. Usage of cell nomenclature in biomedical literature. *BMC Bioinformatics*. 2017;18:17-24.
28. Babic Z, Capes-Davis A, Martone ME, et al. Incidences of problematic cell lines are lower in papers that use RRIDs to identify cell lines. *Elife*. 2019;8:e41676.
29. Witwer KW, Halushka MK. Toward the promise of microRNAs—enhancing reproducibility and rigor in microRNA research. *RNA Biol*. 2016;13:1103-1116.
30. Kilikevicius A, Meister G, Corey DR. Reexamining assumptions about miRNA-guided gene silencing. *Nucleic Acids Res*. 2022;50:617-634.
31. Jacquet K, Vidal-Cruchez O, Rezzonico R, et al. New technologies for improved relevance in miRNA research. *Trends Genet*. 2021;37:1060-1063.
32. Kent OA, McCall MN, Cornish TC, Halushka MK. Lessons from miR-143/145: the importance of cell-type localization of miRNAs. *Nucleic Acids Res*. 2014;42:7528-7538.
33. Chivukula RR, Shi G, Acharya A, et al. An essential mesenchymal function for miR-143/145 in intestinal epithelial regeneration. *Cell*. 2014; 157:1104-1116.
34. Clarivate Analytics Journal Citation Reports. <https://jcr.clarivate.com/>
35. GeneCards. <https://www.genecards.org/>
36. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from micro-RNA sequences to function. *Nucleic Acids Res*. 2019;47:D155-D162.
37. Gopalan S, Devi SL. BNEMiner: mining biomedical literature for extraction of biological target, disease and chemical entities. *Int J Bus Intell Data Min*. 2016;11:190-204.
38. Ye F, Chen C, Qin J, Liu J, Zheng C. Genetic profiling reveals an alarming rate of cross-contamination among human cell lines used in China. *FASEB J*. 2015;29:4268-4272.
39. Huang Y, Liu Y, Zheng C, Shen C. Investigation of cross-contamination and misidentification of 278 widely used tumor cell lines. *PLoS One*. 2017;12:e0170384.
40. Bian X, Yang Z, Feng H, Sun H, Liu Y. A combination of species identification and STR profiling identifies cross-contaminated cells from 482 human tumor cell lines. *Sci Rep*. 2017;7:9774.
41. Gu M, Yang M, He J, et al. A silver lining in cell line authentication: short tandem repeat analysis of 1373 cases in China from 2010 to 2019. *Int J Cancer*. 2022;150:502-508.
42. Boonstra JJ, Van Der Velden AW, Beerens EC, et al. Mistaken identity of widely used esophageal adenocarcinoma cell line TE-7. *Cancer Res*. 2007;67:7996-8001.
43. American Type Culture Collection. <https://www.atcc.org/>
44. Pan H, Bian X, Yang S, He Y, Yang X, Liu Y. The cell line ontology-based representation, integration and analysis of cell lines used in China. *BMC Bioinformatics*. 2019;20:249-258.
45. Patop IL, Kadener S. circRNAs in cancer. *Curr Opin Genet Dev*. 2018; 48:121-127.
46. Rocabado F, Perea M, Duñabeitia JA. Misspelled logotypes: the hidden threat to brand identity. *Sci Rep*. 2023;13:17817.
47. Reid YA. Best practices for naming, receiving, and managing cells in culture. *In Vitro Cell Dev Biol Anim*. 2017;53:761-774.
48. Weiskirchen R. Research reporting guidelines for cell lines: more than just a recommendation. *Annals Trans Med*. 2023;11:421. doi:10.21037/atm-23-1208
49. Byrne JA, Christopher J. Digital magic, or the dark arts of the 21st century—how can journals and peer reviewers detect manuscripts and publications from paper mills? *FEBS Lett*. 2020;594:583-589.
50. Cabanac G, Labbé C, Magazinov A. The ‘Problematic Paper Screener’ automatically selects suspect publications for post-publication (re) assessment. *arXiv*. 2022; (preprint).
51. Hepkema WM, Horbach SPJM, Hoek JM, Halfman W. Misidentified biomedical resources: journal guidelines are not a quick fix. *Int J Cancer*. 2022;150:1233-1243.
52. Wren JD. Clinical trial IDs need to be validated prior to publication because hundreds of invalid National Clinical Trial Identifications are regularly entering MEDLINE. *Clin Trials*. 2017;14:109.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Oste DJ, Pathmendra P, Richardson RAK, et al. Misspellings or “miscellaneous”—Non-verifiable and unknown cell lines in cancer research publications. *Int J Cancer*. 2024;1-12. doi:10.1002/ijc.34995