1   **Evidence provided by high-impact cell culture studies does not support authors' claims**

2

3   Ali Burak Özkaya[1*], Caner Geyik[2,3]

4

5   [1]Department of Medical Biochemistry, Faculty of Medicine, İzmir University of Economics,

6   İzmir/Turkey

7   [2]Department of Medical Biochemistry, Faculty of Medicine, İstinye University, İstanbul/Turkey

8   [3] ISUMKAM Molecular Cancer Research Center, İstinye University, Istanbul/TURKEY

9

10   * ali.ozkaya@ieu.edu.tr

11

12

13

14

15

16

17

18

19

20

21

22

23 **ABSTRACT**

24 **Background**

25 Reliability of preclinical research is of critical concern. Previous studies demonstrate the low

26 reproducibility in research and recommend raising standards to improve reproducibility and

27 robustness. One understudied aspect of this quality issue is the harmony between the

28 hypotheses and the experimental design in published work.

29 **Methods and findings**

30 In this study we focused on highly cited cell culture studies and investigated whether the claims

31 of the study are backed with sufficient experimental evidence or not. We created an open access

32 database containing all 282 claims asserted by 103 different high-impact articles as well as the

33 results of this study. Our findings revealed that only 64% of all claims were sufficiently supported

34 by evidence and there were concerning misinterpretations such as considering the results of

35 tetrazolium salt reduction assays as indicators of cell death or apoptosis.

36 **Conclusions**

37 Our analysis revealed an alarming discordance between the actual experimental findings and the

38 way that the manuscript is written to discuss them in highly cited cell culture studies. In order to

39 improve quality of pre-clinical research, we require a clear nomenclature by which different cell

40 culture claims are distinctively categorized, materials and methods sections to be written more

41 meticulously and cell culture techniques to be selected and utilized more carefully.

42

43 **INTRODUCTION**

44 There is an alarming concern regarding reliability of the published research findings [1]. This is

45 particularly evident in preclinical studies as the clinical translatability is minimal [2]. This low

46 efficiency in research has been discussed extensively in recent years and the lack of

47 reproducibility and overall quality are agreed upon as the main culprits of the problem [3].

48 Reproducibility in preclinical research is estimated to be between 10% to 25% [4,5] and the cost

49   of irreproducible research is calculated to be at least 28 billion USD/year in USA alone [6]. There

50   are many factors contributing to this crisis including; lack of robustness, biased design, use of

51   inadequate models (cell line and/or animal), underpowered studies (insufficient sample size),

52   lack of proper controls (positive, negative), poor use of statistics and the absence of

53   replication/confirmation studies [7]. It is important to note that these design problems often

54   expand to questionable research practices such as p-hacking and cherry picking.

55   Scientists agree that the standards for publishing preclinical research must be raised in a way to

56   encourage robustness and rigor [5,8]. Therefore, many aspects of the preclinical study design

57   have been tacked by various studies over the years. However, the question of whether we can

58   trust results of published preclinical studies remain at large. One aspect that has not been

59   investigated before is the compatibility of the way the manuscript was written with the actual

60   experimental design. More specifically, the relationship between the claims of the studies and

61   the evidence provided to support these claims. In this study, we focused on cell culture research

62   and investigated if the evidence provided by high-impact studies sufficiently supports the claims

63   authors asserted in their manuscript. One of the first things we have noticed during our

64   investigation was the inconsistency in the nomenclature. Many claims such as cytotoxicity,

65   viability, growth, and proliferation were used interchangeably by the authors. Moreover, there

66   were several publications in which only one type of evidence (tetrazolium reduction assay

67   results) was provided to assert various claims. When we searched the literature to refine the

68   consensus nomenclature, much to our disappointment, we could not find any. Many of these

69   terms are not considered uncommon, unfamiliar, or vague enough to be defined in high-impact

70   reviews or guidelines, or to be included in the glossary sections of the molecular biology,

71   biochemistry and even cell culture textbooks. Therefore, we decided to define these terms

72   ourselves mostly based on different sections of "Guidance Document on Good In Vitro Method

73   Practices" by OECD [9] which was the only document we find that might be considered as a

74   consensus nomenclature source. We then carried on our analysis accordingly. As a result, this

75   study contains a nomenclature recommendation as well as the analysis of high-impact cell culture

76   studies.

77   **METHODS**

78   The study consisted of three phases. In phase one, we selected high-impact cell culture studies.

79   In phase two, we identified the claims asserted by the authors as well as the evidence provided

80   by them to support these claims. In the final phase we analyzed sufficiency of the evidence for

81   each of the claims.

82   **Article Selection**

83   We searched Web of Science (WOS) database (Clarivate Analytics) for studies that contain at least

84   one of these keywords: "cytotoxicity, viability, cell death, growth inhibition, proliferation, or anti-

85   cancer". We included original research articles using *in vitro* techniques. The search string below

86   was used in advanced search feature of WOS:

87   WOS core database (TS= ("cytotoxicity" OR "viability" OR "cell death" OR "growth

88   inhibition" OR "growth inhibitory" OR "proliferation" OR "anti cancer") AND TS=("cell

89   culture" OR "in vitro" OR "cell line")) AND LANGUAGE:(English) AND DOCUMENT TYPES:

90   (Article)

91   Studies published in 2017 and 2018 were retrieved in 10.03.2020. We exported the data as an

92   excel file, sorted the list based on citations received and selected the most cited 121 publications

93   (each receiving at least 65 citations) as high-impact studies for further analysis. Upon

94   investigation of the articles, we excluded 18 studies that do not contain a claim, were not carried

95   out in cell culture, or were not original works. After exclusion of these ineligible studies, 103

96   studies were left for claim analysis.

97   **Claim Selection and Definitions**

98   We identified seven distinct claims including: anti-cell, apoptosis, cell death, cell growth,

99   cytotoxicity, proliferation, and viability in the studies. Using any of these terms to present or

100   discuss a finding was considered as a claim.

101   We outlined our definitions for these claims as follows by using OECD guide GIVIMP [9]:

102   *Proliferation rate* represents how fast a group of cells divide over time. If the measurement was

103   an end-point measurement, then it must be able to differentiate the decrease in division

104   capabilities of the cells from cell death to provide sufficient evidence for proliferation rate. The

105   reason is simple: if the treatment of question induces cell death in the treatment group, there

106   would be fewer living cells (compared to untreated control) without a decrease in proliferation

107   rate. Methods specifically focusing on replication rate (such as nucleotide incorporation) or

108   measuring the number of viable cells without a treatment over time as well as real-time

109   observations and proliferation markers were considered as sufficient evidence.

110   *Viability* represents the number of living cells. It is the broadest term since there is no

111   specification regarding the factor affecting the number (such as proliferation rate or cell death).

112   Any method directly measuring the number of living cells and methods measuring metabolic

113   activity were considered as sufficient evidence for viability.

114   Since *cell death*, by definition, requires cells to die, end point analysis measuring the abundance

115   of living cells cannot provide sufficient evidence for this claim as the measurement does not

116   differentiate the decrease in number due to dying cells from slowed-down proliferation rate. Any

117   assay measuring death-related alterations such as membrane integrity or cell-death specific

118   markers was considered to provide sufficient evidence.

119   *Apoptosis* is a form of programmed cell death which has well established and characterized

120   distinct features. The methods that are capable of demonstrating the changes in these features

121   such as phosphatidylserine exposure on the outer membrane, DNA fragmentation,

122   morphological changes and molecular switch responsible in apoptosis were considered as

123   sufficient evidence to show apoptosis.

124   *Cytotoxicity* indicates being toxic to cells. Being toxic itself is a broad term and there are

125   conflicting definitions in use. We decided to consider it as cell death instead of decreased viability

126   as the most widely accepted capability of a toxic agent is killing (as in cytotoxic T cells and

127   cytotoxic chemotherapy), and accepted evidence indicating cell death as sufficient to prove

128   cytotoxicity. This decision had an impact on the final analysis which might be considered

129   controversial as many of the articles might have used the term to represent viability decrease.

130   We addressed this in results section. We decided to consider *anti-cell* (which was asserted just

131   once) as if it was referring to cell death as well.

132 *Cell growth* may indicate either proliferation rate or the size change of cells depending on the

133 definition embraced. Since, there already is a term representing proliferation rate as the name

134 implies, we first considered to accept it as a measure of increased cell size. However, after looking

135 at the articles in our list, we realized that the term almost exclusively used to indicate

136 proliferation rate and consequently we embraced that definition in our analysis.

137 **Database Construction**

138 We constructed a database in Airtable to carry out evidence analysis. Information from WOS

139 database such as "article name", "DOI", "citation count", "journal name" as well as our

140 parameters of interest such as "claim", "evidence", "method", "sufficiency of evidence", and

141 "subject area" were entered for every article investigated.

142 Here, "method" represents scientific methods used in the study whereas "evidence" is defined

143 as a supergroup of methods measuring same biological phenomenon. For example, two separate

144 methods such as lactate dehydrogenase (LDH) activity assay and PI both of which measure

145 membrane damage as an indicator of cell death were classified into "membrane integrity"

146 evidence supergroup. Similarly, various tetrazolium and resazurin reduction assays were

147 considered to provide "dehydrogenase activity" evidence which is an indicator of cellular

148 metabolic activity.

149 We have also divided the studies in two notional groups of "subject area" first being

150 "Biochemistry, Molecular Biology, Genetics, and Medicine" and second "Chemistry,

151 Chemical/Biomedical Engineering, and Materials Science" based on field information provided

152 by WOS.

153 Database is accessible via the link: https://airtable.com/shrClTE87e1l28ExG

154 **Evidence Analysis**

155 Evidence was analyzed for each of the claims by a case-by-case approach. We refer to the

156 definitions we have embraced and the OECD guide GIVIMP [9] to determine if the measured

157 parameter provides sufficient evidence for that claim. Table 1 summarizes the detected claims

158 and the evidence considered as sufficient or insufficient. It is important to note that if there were

159 multiple evidence for a single claim, we focused on the strongest evidence as well as the strength

160 of the combination of the evidence. Therefore, in the table if the evidence is listed under

161 "insufficient" it means it does not provide sufficient evidence by itself. For example, even though

162 expression of Blc-2 members (such as Bax/Bcl-2 ratio) is commonly investigated along with other

163 markers of apoptosis, it is classified as insufficient because it does not provide proper evidence

164 for apoptosis by itself.

165 There were some cases where the same type of evidence may be sufficient or insufficient based

166 on the experimental design. Viability indicators such as cell count, colony formation, DNA

167 amount, and dehydrogenase activity were considered to provide sufficient evidence for

168 proliferation only if there is no treatment (hence no reason for a change in cell death ratio) and

169 the study is designed in a way to observe proliferation of these untreated cells. Moreover, if the

170 method utilized to obtain the evidence was not mentioned in materials and methods section of

171 the article, we then classified it as insufficient (independent from the actual evidence) which was

172 the case for cell count providing insufficient evidence for viability and proliferation (article id#7

173 from Airtable database). Another such deviation was the membrane integrity as an evidence for

174 viability. Cell death marker PI was used as an evidence for viability in one article (article id#66

175 from Airtable database) which was considered as insufficient. All other articles (6 articles in total)

176 claimed viability by providing different membrane integrity assays (including 7-AAD, GhostDye,

177 Trypan Blue and Zombie UV) all of which measure viable cells.

178 **Table 1**. List of sufficient and insufficient evidence for claims.

Claim	Sufficient	Insufficient
Anti-cell		Dehydrogenase activity
Apoptosis	Caspase-3 activity, Cell morphology, DNA fragmentation, Phosphatidylserine exposure	Dehydrogenase activity, Membrane integrity, Bcl-2 Family expression
Cell Death	Membrane integrity, Phosphatidylserine exposure	Cell count, Dehydrogenase activity, Esterase activity
Cell Growth	Cell count, Colony formation, Nucleotide incorporation	Dehydrogenase activity
Cytotoxicity	Membrane integrity, Phosphatidylserine exposure,	Dehydrogenase activity

Proliferation	**Cell count***, **Colony formation***, **Dehydrogenase activity*, DNA amount***, Dye inclusion, Ki67 expression, Nucleotide incorporation, Phospho-histone H3	Actin, **Cell count***, **Colony formation***, **Dehydrogenase activity*, DNA amount***, Esterase activity, Signaling Pathways
Viability	ATP amount, **Cell count***, Dehydrogenase activity, Dye inclusion, Esterase activity, **Membrane integrity***, Total protein	**Cell count***, **Membrane integrity***, Signaling Pathways, Unknown

179   * The evidence is either sufficient or insufficient for given claim depending on the study design.

180   **The Effects of the Journal and Subject Area on Evidence Sufficiency**

181   After the sufficiency of evidence was assessed for each claim, we then sought to find whether

182   being published in a specific journal or in a subject area would affect sufficiency rate. Seven

183   journals were selected for analysis as they meet our criteria of having at least 10 claims, namely

184   *Biomaterials, Cell, Cell Death Dis., Mol. Cancer, Nature, Oncotarget, Ann. Biomed. Eng.* (97 claims

185   out of 282, 34.40%). Every journal was compared to the complete data set excluding themselves.

186   Two subject areas: "Biochemistry, Molecular Biology, Genetics, and Medicine" and "Chemistry,

187   Chemical/Biomedical Engineering, and Materials Science" were compared to each other. Fisher's

188   exact test (2-tail) were used in all comparisons.

189   **RESULTS**

190   We have investigated 282 claims asserted by 103 different high-impact articles. We identified 7

191   unique claims supported by 20 types of evidence all of which was obtained via 40 different

192   methods. Details are presented in the database (https://airtable.com/shrClTE87e1l28ExG)

193   The most common claim was proliferation rate changes with 85 claims, followed by viability,

194   apoptosis, cytotoxicity, cell death, growth changes and anti-cell activity (Figure 1A). Upon

195   investigation, we considered the evidence of 102 claims (36%), which was asserted by 66

196   different studies (64%), as insufficient (Figure 1B). Claims of cytotoxicity (11 sufficiency rate in 37

197   claims, 30%), proliferation (35 in 85, 41%), and cell growth (10 in 18 claims, 56%) particularly

198   lacked proper evidence. Viability (66 in 72, 92%), apoptosis (46 in 51, 90%), and cell death (14 in

199   18, 78%), on the other hand, were more frequently claimed with proper evidence (Figure 1A).
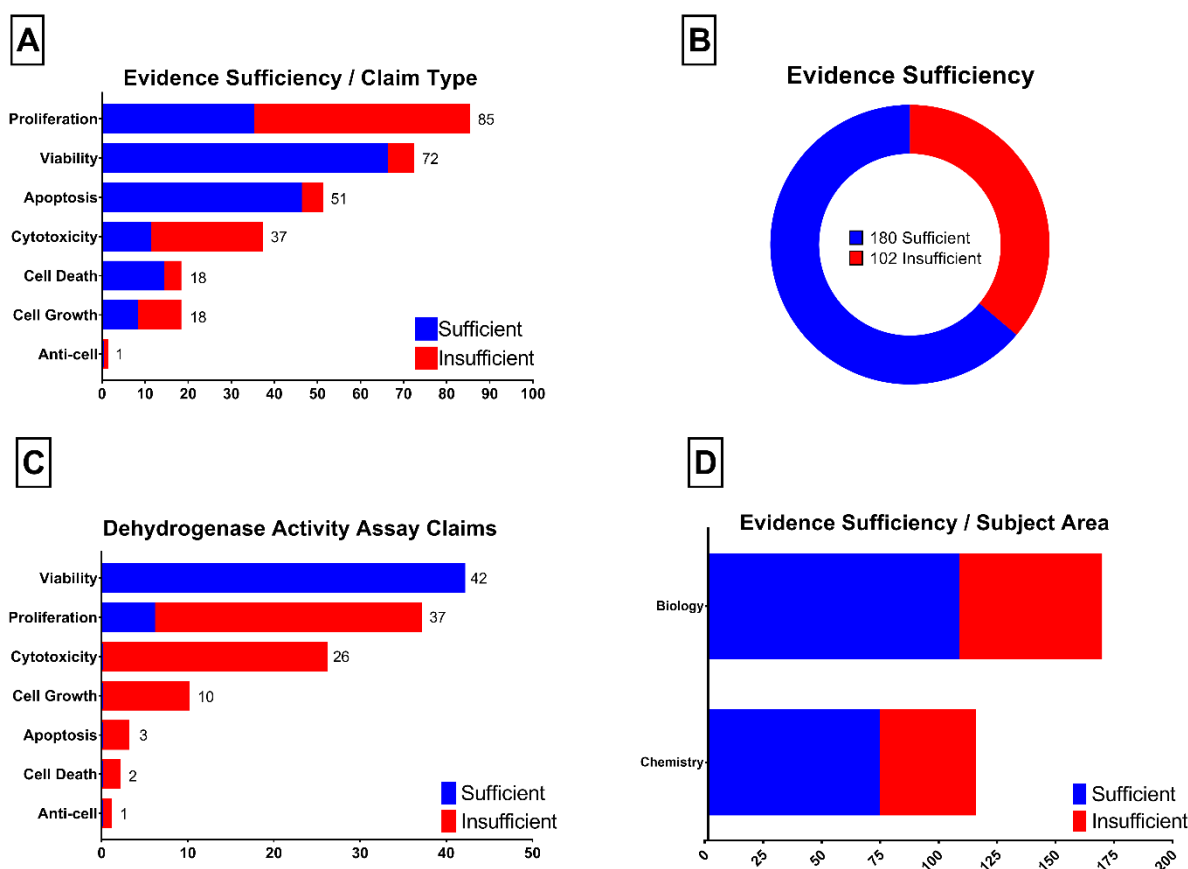
Figure 1. Evidence sufficiency analysis results. 1A. Evidence sufficiency according to the claim type; 1B. The ratio of claims that are supported by sufficient evidence or not; 1C. Evidence sufficiency of claims supported by dehydrogenase activity assay; 1D. Evidence sufficiency according to the article subject area

When evaluating claims, we had to make assumptions on behalf of the researchers by assigning specific definitions to the claims which is listed in the materials and methods section of this manuscript. Cytotoxicity was most likely to be used differently in articles we analyzed, possibly affecting the results of our analysis. If we had considered cytotoxicity as a measure of viability instead of cell death, all the cytotoxicity claims would have had sufficient evidence increasing the overall sufficiency rate from 64% to 73%.

Measurement of dehydrogenase activity was by far the most common type of evidence. 121 different claims (44% of all claims) including viability, proliferation, cytotoxicity, cell growth, cell death, anti-cell and even apoptosis put forward dehydrogenase activity findings as evidence

215   (Figure 1D). The most common method used to measure dehydrogenase activity was tetrazolium
216   reduction assay supporting 115 different claims (41% of all claims). Measurement of
217   dehydrogenase activity was followed by measurement of phosphatidylserine exposure (31
218   claims) and membrane integrity (29 claims).

219   While the results of dehydrogenase activity assays were interpreted correctly only in 73 claims
220   (73 in 121, 60%), the second most common evidence, measurement of phosphatidylserine
221   exposure, was interpreted correctly (as an indicator of apoptosis) in all related claims (31 in 31,
222   100%). Similarly, measurement of membrane integrity was mostly correctly utilized as evidence
223   for cell death claims (27 in 29, 93.1%) (Figure 1C)

224   We also analyzed whether the subject area or the journal that the article was published in might
225   be an indicator of evidence claim relationship. According to our analysis, the subject area did not
226   have a significant influence over evidence sufficiency (Figure 1D). Similarly, amongst the most
227   frequently appeared journals in our database none of them demonstrated a significant difference
228   when compared the overall claims supported by sufficient evidence.

229   **DISCUSSION**

230   Our findings reveal a discordance between the claims and the evidence of high-impact cell culture
231   studies. This is especially evident in studies utilizing the findings of tetrazolium reduction assay
232   alone to support various claims. Striking examples include article id#9 claiming viability,
233   proliferation, cytotoxicity and apoptosis changes, and article id#26 claiming viability,
234   proliferation, growth changes and anti-cell activity with results from this assay. This is partly
235   because these assay kits are advertised by their manufacturers as a tool to measure viability,
236   cytotoxicity, proliferation and growth. Combining this with being relatively easier to perform and
237   affordable leads to these assays being perceived as a one-size-fits-all solution by research groups
238   wishing to avoid more complicated cell culture techniques. However, this reductionist approach
239   makes it difficult for the findings obtained from the study to provide meaningful answers to the
240   research questions of the article. Even though the reduction of cellular metabolic activity is not
241   a clear indicator of apoptosis or cell death, since that statement exists in a high-impact article,
242   there is a good chance that it will be cited as such as well.

243   In fact, the articles we analyzed were cited more than 9000 times as of March 2021 (within two

244   to four years). Even though the claims without sufficient evidence may not be the reason for

245   citation in most of the cases, the impact of such studies with unreliable findings on preclinical

246   research is undeniably large. Many agree that we need strategies in place to improve the

247   standards for pre-clinical research [5,8]. As a part of this process, we need clear and distinct

248   definitions for terms corresponding to the claims asserted in the articles. In this work we offered

249   a nomenclature recommendation by which the most common claims in cell culture studies may

250   be distinctively expressed. Moreover, scientists both as authors and reviewers are required to

251   ensure that the experimental designs reflect the research questions asked. As the findings of our

252   study indicate, we believe a more meticulously written materials and methods section and

253   careful selection and utilization of cell culture techniques are critical to raise robustness and

254   overall quality of preclinical research.

255   **ACKNOWLEDGEMENTS**

257   **COMPETING INTERESTS**

258   Authors declare no competing interest.

259   **REFERENCES**

260   1.   Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Med. 2005;2: e124.
261   doi:10.1371/journal.pmed.0020124

262   2.   Takebe T, Imai R, Ono S. The Current Status of Drug Discovery and Development as
263   Originated in United States Academia: The Influence of Industrial and Academic
264   Collaboration on Drug Discovery and Development. Clin Transl Sci. 2018;11: 597–606.
265   doi:10.1111/cts.12577

266   3.   Peers IS, Ceuppens PR, Harbron C. In search of preclinical robustness. Nat Rev Drug Discov.
267   2012;11: 733–734. doi:10.1038/nrd3849

268   4.   Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data
269   on potential drug targets? Nat Rev Drug Discov. 2011;10: 712–712. doi:10.1038/nrd3439-c1

270   5.   Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012;483: 531–
271   533. doi:10.1038/483531a

272   6.   Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical
273   Research. PLOS Biol. 2015;13: e1002165. doi:10.1371/journal.pbio.1002165

274   7.   Ioannidis JPA, Kim BYS, Trounson A. How to design preclinical studies in nanomedicine and
275   cell therapy to maximize the prospects of clinical translation. Nat Biomed Eng. 2018;2: 797–
276   809. doi:10.1038/s41551-018-0314-y

277   8.   Begley CG, Ioannidis JPA. Reproducibility in science: improving the standard for basic and
278   preclinical research. Circ Res. 2015;116: 116–126. doi:10.1161/CIRCRESAHA.114.303819

279   9.   OECD. Guidance Document on Good In Vitro Method Practices (GIVIMP). 2018. Available:
280   https://www.oecd-ilibrary.org/content/publication/9789264304796-en

281