**DELPHI ROUND 1 – REPORT**

**DEVELOPMENT OF A RoB TOOL FOR DENTAL MATERIALS STUDIES**

This DELPHI Round gathered 26 voters, which were either key experts in the field **(16%)** or belonged to the IADR-DMG **(40%)** or EFCD Board **(44%)**. A list of the participants is shown below (Table 1).

**Table 1.** Participants who took part in the 1st DELPHI round for RoB development.

|  |
| --- |
|  |
| 1 | **Antonio Delgado** |
| 2 | **Salvatore Sauro** |
| 3 | **Sebnem Turkun** |
| 4 | **Jack Ferracane** |
| 5 | **Alessandro Loguercio** |
| 6 | **Klaus Neuhaus** |
| 7 | **Alvaro Della Bona** |
| 8 | **James Tsoi** |
| 9 | **Will Palin** |
| 10 | **Satoshi Yamaguchi** |
| 11 | **Adriano Lima** |
| 12 | **Niek Opdam** |
| 13 | **Fabricio Collares** |
| 14 | **Lorenzo Breschi** |
| 15 | **Mattias Hannig** |
| 16 | **Sophie Domejean** |
| 17 | **Frode Staxrud** |
| 18 | **Laura Ceballos** |
| 19 | **Nicola Scotti** |
| 20 | **Julia Amato** |
| 21 | **Sebastian Paris** |
| 22 | **Mutlu Ozcan** |
| 23 | **Annalisa Mazzoni** |
| 24 | **Vinicius Rosa** |
| 25 | **Falk Schwendicke** |
| 26 | **Mary Anne Melo** |

**Domain 1: Bias in Planning and Allocation**

***Q1.1***

*“Do you agree with the first source of bias identified within this domain, which is "Control group featured in the experimental design"?”*

**92.3% of the participants agreed with Q1.1**

**Comments:**

Control groups are critical to any meaningful experimental design. A true "control" is often difficult to identify with, say, commercial material testing since the constituents are usually not fully known.

Will the RoB stipulate how should be an adequate control group? For example, in composite repair evaluation, the study must contain one control without surface treatment and adhesive and a control evaluating the cohesive strength of the resin for a fair assessment. RoB should guide the authors to observe when the study needs one or more control groups. Is there any way to drive the authors to find the best control group for different studies?

Especially for studies on adhesives and composites it would be good to define some brands as 'gold standard' to be used as control group, a.g. Optibond FL, SE Bond. These materials should have a robust body of evidence. In my opinion, it is more valuable to include SE Bond as 2-step (mild) self-etching as control, then 'any 2-step self-etching adhesive"

Not all studies in dental materials will define a "positive or negative" control. Many of them use materials as benchmark. One cannot compare the longevity of composites to "negative control/no restoration" but maybe with amalgam (benchmark). The positive/negative control is required but provision needs to be done for benchmarking materials and "anticipated outcomes".

***Q1.2***

*“Do you agree with the final source of bias identified within this domain, which is "Sample size calculation"?”*

**80.2% of the participants agreed with Q1.2**

**Comments:**

Sample size calculation, power tests, etc are critical for experiments with high levels of variability (for populations, clinical data, etc). This is not the case for well controlled in vitro (non-biological, material-based) experiments where less variability is likely.

Too many times sample size is only based on previous similar performed studies, which does not mean the sample size would be effective in testing the initial hypothesis

The RoB should specify, if possible, the parameters that should be described by the authors of the studies to rank the description of sample size calculation as adequate or if it has missing information.

Not all sample sizes are calculated. Whether we like or not, sample size in dental materials research not calculated but often defined by current literature/standards.

Sample size availability seem to be equally/more relevant than sample size calculation.

How to calculate the sample size is critical.

Increasing the sample size tends to reduce the sampling error; that is, it makes the sample statistic less variable. However, increasing sample size does not affect survey bias.

Iso standards are an exception and some characterization methods (i.e., mass spectroscopy)

Whenever sample size is not calculated, it needs to be assumed based on literature. Yet, once you do that, it would be easy to calculate the sample size...

***Q1.3***

*“Do you agree with the final source of bias identified within this domain, which is "Correct randomization of samples"?”*

**84.6% of the participants agreed with Q1.2**

**Comments:**

How to avoid the bias within randomization in the samples would be important.

Yes, and for teeth used for example for bond strength testing it is important to understand statistical degrees of freedom, i.e., results are likely to be highly correlated for one tooth. Thus, the statistical degrees of freedom might only be 1 per tooth, instead of the multiple test pieces that are assumed for collection of sticks from it, i.e. the statistical variance within and between teeth are important.

Again, the RoB will probably help the authors with the term "adequate", giving examples of a sound method or a low accurate method.

This is very vague. Does that mean that is a bovine tooth is used, it should be from different teeth (different cows?). and if we adhere to ceramic? different ceramic samples, or the same?

Strongly agree in methods with teeth.

What is correct? What is adequate? These terms need to be defined in order to assess if randomization was "correctly" performed. Randomly allocated (where the researcher sorts "this goes here, that goes there) is different from allocated via randomization via lists created by software’s. It is important to define what is an "adequate randomization method" first.

**Domain 2: Bias in Specimen Preparation**

***Q2.1***

*“Do you agree with the first source of bias to be included in this domain: "Identical experimental conditions across groups"?”*

**88.4% of the participants agreed with Q2.1**

**Comments:**

The meaning of "identical experimental conditions" should be defined. For example, in mechanical testing, same equipments, jigs, and so on operated in the same lab might have yielded some different results - which is indeed the compliance of the machine and the jigs matter. I would suggest method validation (like stated in ICH) can be a way out rather than setting "identical experimental conditions". In addition, if the samples (say surface treatment) are not the same so how can we claim "identical" ?

It could be added to the description if the groups were identified or blinded to the person who prepared the specimens. The blind preparation can reduce the bias since any group cannot be favoured, even unconsciously.

Sample storage is a key factor. When working with extracted teeth, the time from extraction, the storage immediately after extraction and the sample preparation time could be a strong bias in lab studies.

***Q2.2***

*“Do you agree with the second source of bias to be included in this domain: "Standardization of samples and materials"?”*

**88.4% of the participants agreed with Q2.2**

**Comments:**

It should be mentioned if human or bovine teeth were used. In human teeth, there might be a high variability of hard substance quality and properties, depending on the tooth age and its service time in the mouth. Probably, the sample size should be larger in human teeth than in bovine teeth.

Some key parameters of the samples and materials should be reported and controlled. These parameters should be easy to read and repeatable with other group.

This also relates to the degree of freedom comments in my previous answer. Also, accurate reporting of material composition of commercial materials is impossible as manufacturers do not divulge complete information.

I personally would recommend writing always the step-by-step procedures followed to apply a material instead of writing "following the manufacturer instructions".

Can be improved. It is easier to "standardized samples and materials" when comparing commercially available materials. However, research where new materials/applications are "developed", this is not easily reported. I had an issue in the past where my paper was assigned as "high bias" because "histology was not evaluated by independent assessors". You see, in clinical trials data is evaluated independently, in reality (histology) you don't engage three people to check slides... Same goes here, this second "source of bias" will create a space for experimental papers to be judged as "high bias" because the procedures for novel materials cannot be equated to commercial ones. When people read the research paper, they understand the particularities. But once a "standard for bias classification" is set, then the paper is "high risk of bias" because the context is removed once the classification is set. It seems that this risk applies to "comparison of commercially available materials" only

**Domain 3: Bias in Outcome Assessment**

***Q3.1***

*“Do you agree with the first source of bias to be included in this domain: "Standardization of the testing procedures and outcomes"?”*

**88.4% of the participants agreed with Q3.1**

**Comments:**

We may be careful on this. Innovation is usually association with new approaches and procedures, therefore, it should be allowed some flexibility to improve existing ("standards") procedures

Even the experiment claims to follow certain standards or guidelines, there is no guarantee whether the results are coming out with good quality , i.e. reliable. Many universities are running research labs but not certified testing labs that have periodic calibration, method validations and QMS. Following the procedure without validate the methods (as in ICH) in the operating equipment gives no sense.

Sometimes, methods are not in ISO or ADM standards

For better generalizability ADM and ISO should only be mentioned as examples of guidances and standards (just wording).

ADM guidelines are not prescriptions and ISO standards are not made for research purpose. the recommendation for following these documents needs to be carefully reconsidered.

***Q3.2***

*“Do you agree that the second source of bias to be included in this domain should be: "Blinding of the testing operator"?*

**84.6% of the participants agreed with Q3.2**

**Comments:**

I think this is good, but rarely will be followed - just not really practical

This is simply difficult to do.

Depends on the outcome. When based on evaluation scales that are evaluator dependent yes (like level of leakage), less important when a machine determines the outcome

Whenever it is possible

Often not feasible in a practical situation

I strongly disagree/oppose with this. How can researchers around the world blind their PhD students from the work they are doing? This statement is very dangerous because every paper will be judged as "high bias". People manufacturing the samples are testing samples and that's life and all is okay. It is nearly impossible to have one person manufacturing samples while other test independently. I have done this for a clinical study and it's a nightmare, schedules do not match, cost increase, samples are spoiled. In an ideal situation, yes, it will be perfect, in reality, this will make every paper arising from any PhD thesis to be classified as high bias. The classification cannot be detached from reality. I highly oppose this statement.

**Domain 4: Bias in Data Treatment and Outcome Reporting**

***Q4.1***

*“Do you agree with the second source of bias to be included in this domain: "Standardization of samples and materials"?”*

**92.3% of the participants agreed with Q4.1**

**Comments:**

Statistics need to be reviewed. In lab test, it should be more controlled and the statistical

analysis should not be complex.

Again, my previous comments on degrees of statistical freedom are also relevant here.

Mandatory

To the best of their knowledge. Rephrase. Choice and description of stastical tests are right or wrong. There is no "best of knowledge". Either the right or wrong test can be applied.

***Q4.2***

“Do you agree with the second source of bias to be included in this domain: "Correct reporting of outcomes"?”

**92.3% of the participants agreed with Q4.2**

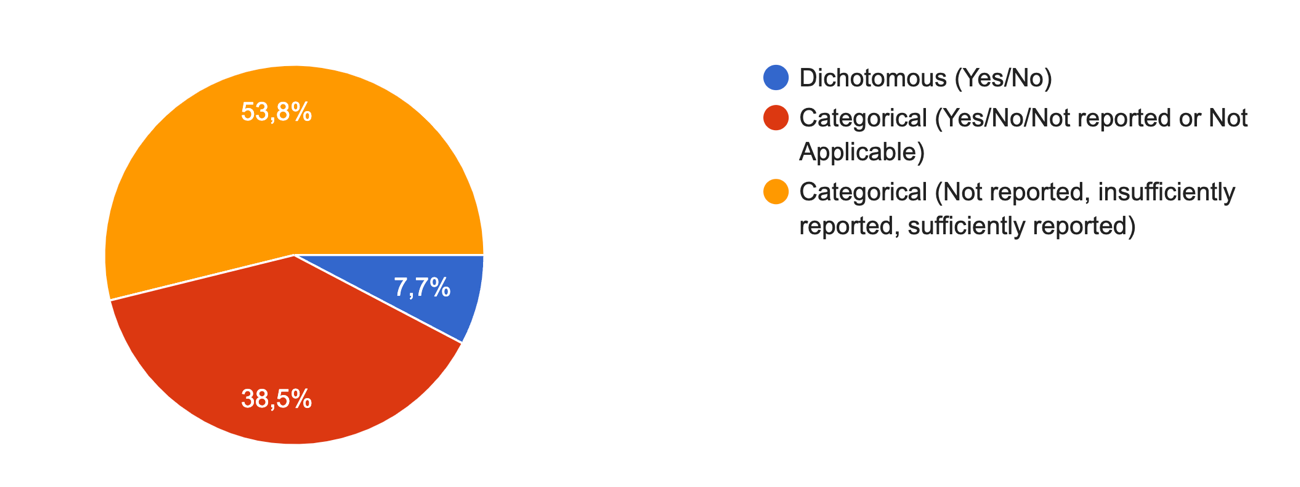
**Comments:**

In most studies, qualitative data (i.e. fractography observations) is essential to support the quantitative values.

It can be added to the description of D4.2 if the authors cite differences without statistical significance, only based on the numerical difference between the groups. This fact described can lead the readers to equivocated conclusions.

1) Not all data is evaluated for all groups in every study. In several occasions, testing parameters are dropped out for some groups because they do not make sense to be tested. 2) (i.e. are failure modes and fractographic analysis missing in bond strength studies?) : this goes against academic freedom. If the researcher has opted not to evaluate failure modes, it does not mean that the study is biased. I work a lot with stem cell differentiation and it is widely known that genetic expression just tells half of the story, protein analyses are always required. However, i cannot say that a paper with gene expression is biased because it is not perfect. Also, if gene expression is similar, no need to measure protein in many cases... Hence, is there bias? No. My impression is that this statement is recommending "what should be evaluated" and this will create a feeling of "bias" disconsidering what the authors had in mind. Can we say on fracture strength of cermisn that a paper IS BIASED because not fractographic analysis is avaiable? probably not...

**Types of Response Choices**

****

**Figure 1.** Response choice distribution for the answer scale to be used with the new RoB tool.

**Comments:**

By doing this way, we can give more feedbacks to the authors

I like the last one, but i think it might make sense to include a not applicable as well

I believe that in some cases, not applicable is a possible answer (i.e. operator blinding when two different material will be applied)

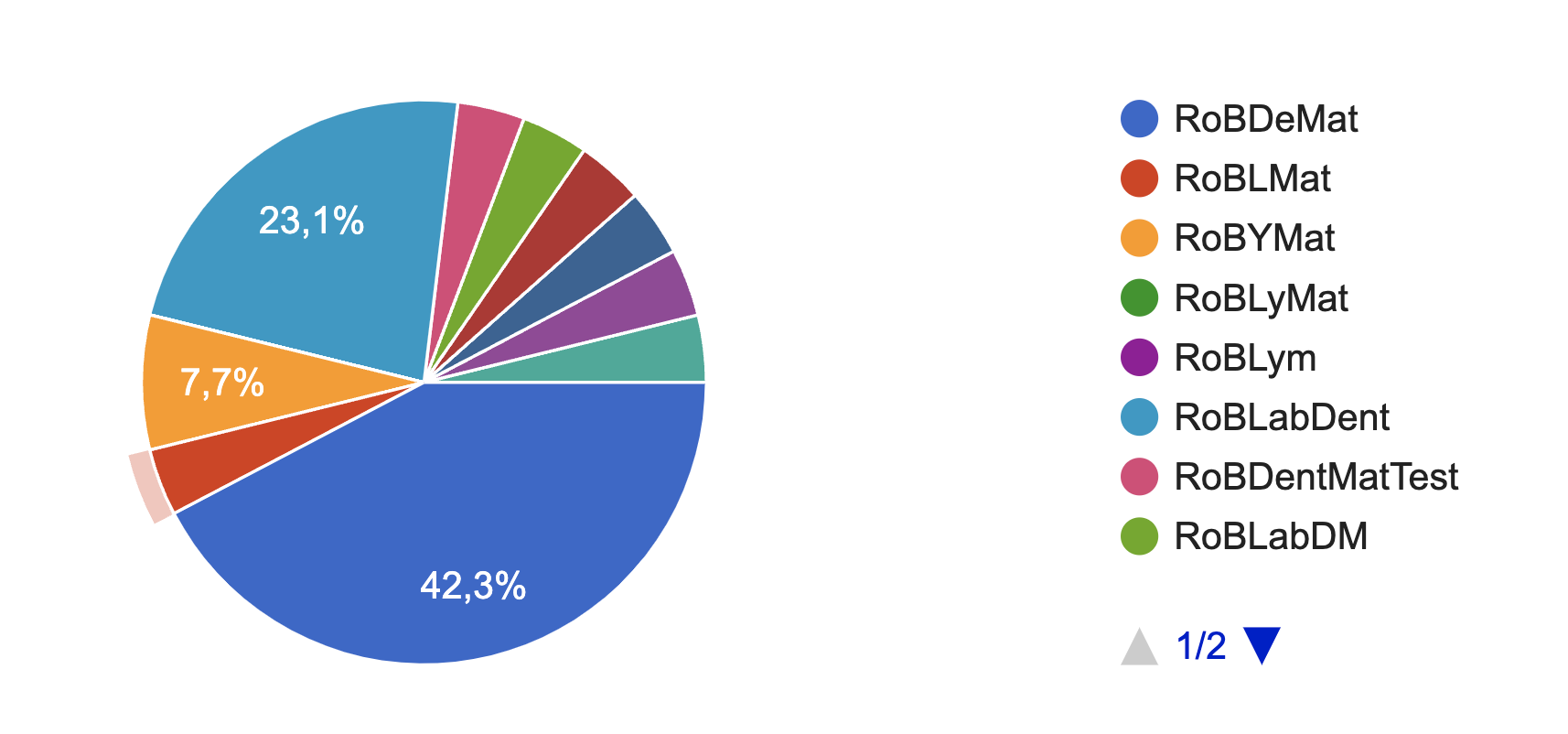
Should also contain Not applicable

Not reported and not applicable are not the same, and they should be distinguished for further analysis.

Yes/no does not take in consideration the various nuances present in R&D research

**Types of Response Choices**

The two most voted acronyms were **RoBDeMat (42.3%)** and **RoBLabDent (23.1%).**



**Figure 2.** Voting distribution for the acronym of the RoB Tool.