PEER REVIEWED

Optimising Endpoints and Educating Investigators to Drive Clinical Trial Success

Everything in a clinical trial should be designed to optimise the success of meeting the trial's endpoints. This goal starts with developing a flawless protocol, which then dictates all downstream trial activities. Trial Sponsors often in collaboration with Contract Research Organisations (CROs) then painstakingly follow this protocol, which was optimised in every way to meet the trial's endpoints, yet most clinical products eventually fail to make it to market.[1] Failing to meet the trial's primary endpoint could mean that the treatment does not work as hypothesised, but trial endpoint failure could also stem from several other factors such as not optimising your protocol's inclusion and exclusion criteria, from treatment non-adherence, or from a highly variable measurement.

Measurement variability could stem from the instrument (e.g., measurement repeatability), biological variability (e.g., inherent differences among participants such as height), environmental conditions (e.g., temperature may affect some measurement devices), or simply not following the protocol as intended.[2] While most Investigators intend to follow the protocol, deviating from the protocol could stem from lack of clarity in the written trial documents, method complexity, participant cooperation, or biases based upon the Investigator's training history (2). While it is challenging to fully control for all these issues, most of these problems can be mitigated with thoughtful methodological implementation and comprehensive multi-modal training.[2,3] This article explores strategies for optimising clinical trial endpoint via optimising the methods used to collect the endpoints data and by training the investigator to correctly complete these measures. This article will accomplish this goal by first describing three common, yet diverse endpoints used in ophthalmology trials to illustrative examples of how to control for every endpoint factor possible and it will then discuss investigator endpoint training and why it is so important to control for all these factors from a statistical standpoint.

Symptomatology

Understanding patient symptoms is one of the most basic aspects of patient care in any clinical discipline. An evaluation of symptomatology in clinic typically starts with the clinicians simply asking the patient how they are doing. While these conversations can be powerful for diagnosing a patient, these conversations can easily take multiple directions and they have limited applicability to the research setting due to the challenges of standardising open-ended conversations. If we use dry eye disease as an example, subjective symptoms are a key competent of diagnosis,[4] and in the clinical setting, patients may specifically report symptoms such as dryness, soreness, burning, blur, light sensitivity and itch,[5] which can be directly queried by the clinician. Standardising this line of questioning begins with surveys. While there are several standardised dry eye surveys,[4] one of earliest, most common and simple ones in this indication and other indications is the visual analog scale (VAS).[6] A VAS consists of a horizontal line (continuous scale without numbers) with an anchor on each end; the left-side anchor typically represents the complete absence of symptoms while the right-side anchor typically represents the most extreme presentation of symptoms (Figure 1).[6–8] When a VAS is administered, the participant is asked to place a mark on the line to represent their current symptoms state.[3–5] This line, which is usually 100 mm long, is then converted to a symptom score by dividing the distance between the left anchor and the mark by the total length of the line and multiplying this decimal by 100 to obtain a symptoms score.[3–5] A VAS has the advantage of having a high degree of sensitivity since patients can place a mark anywhere on the continuous horizontal line and because of this they are well suited for measuring changes within a participant over time.[8] This test is enhanced by clearly educating the participant on the directions, the investigator on how to score the metric and the conditions themselves such as having a printing company print all of the VAS tests for a trial and printing them in a way that produces in a line that is exactly 100 mm long. While the above noted standardisations efforts have dramatically improved the repeatability of this metric, the dry eye community has further developed this metric with psychometric statistical approaches and created a further standardised survey called the Symptom Assessment iN Dry Eye (SANDE).[9] The SANDE is a 2-item VAS used to quantify the frequency (question 1) and severity (question 2) of dry eye and/or ocular irritation symptoms, and it uses a global score to evaluate symptoms severity and to track treatment effectiveness across a trial.[9]

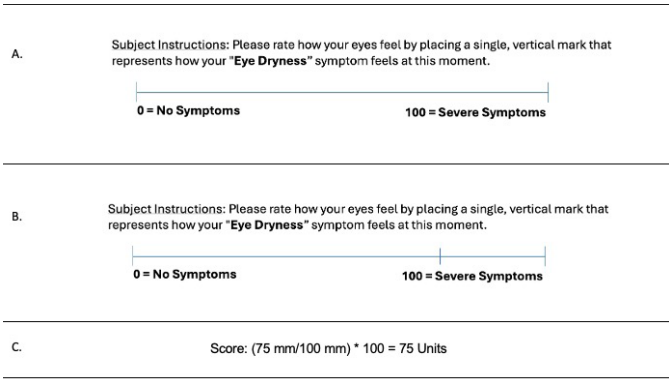

Figure 1: Eye Dryness Visual Analog Scale

Best Corrected Visual Acuity

Visual acuity is one of the most if not the most iconic eye test.[10] Patients report to clinic and they expect to read an eye chart to determine their visual potential and if they need to wear spectacles or if they need a change in their spectacle prescription. One of the first standardised visual acuity charts is the Snellen chart, which was invented by Herman Snellen in 1862. The Snellen chart is the classic eye chart that has the large "E" at the top of it.[10] While the Snellen chart can adequately determine visual acuity in most clinical situations, it has several shortcomings in the research space. Snellen charts specifically have an unequal number of letters on each line and it has unequal visual angle changes between lines, which effectively makes them more of a categorial measurement (placing participants into visual acuity bins).[10] The research shortcomings of Snellen charts were mitigated by the advent of logarithm of the minimum angle of resolution (logMAR) charts.[11–13] Bailey and Lovie (Bailey-Lovie chart) introduced the first logMAR chart back in 1976.[10] The magic of logMAR charts is that they are highly standard measures that systematically evaluate visual acuity letter-by-letter until the participant's visual acuity is determined.[14] Each line on the chart is made up of 5 standardised letters.[10] The line on the top has the largest letters, and the letters get smaller as one reads down the chart.

Each line gets progressively smaller by exactly 0.10 logMAR and each letter on a line is worth 0.02 logMAR units. Thus, logMAR charts can be analysed linearly like incremental changes on a line. This is important because this approach increases discernability between two measurements and it makes it easier to determining a statistical difference between groups.

The gold standard for evaluating visual acuity in any clinical trial is the Early Treatment of Diabetic Retinopathy Study (ETDRS) charts.[14] The ETDRS chart originates from the National Institute of Health-funded Early Treatment of Diabetic Retinopathy Study.[14] The ETDRS chart was introduced in 1982 and it is now distributed by multiple manufacturers.[10] The distributed chart is a set of three standardised charts. Chart R is used for refraction (determining spectacle prescription), Chart 1 is used for determining right eye visual acuity and Chart 2 is used for determining left eye visual acuity. This approach was adopted to avoid memorisation during testing (e.g., avoid recall bias while testing left eye). The logMAR design has since been applied to other types of visual acuity charts (e.g., charts adapted for measuring vision in children). Other testing safeguards that have been implemented include scoring sheets that include every letter on the chart, which is marked by the testers as read or not read during the test to help ensure recording accuracy. Methodological expertise is furthermore often supported by testers being required to read an in-depth training manual, completing a written knowledge test and by the tester completing a in person or virtual practical to determine if the tester is following the protocol. The overall evolution of visual acuity testing has been tremendous and it has evolved into one of the most frequent primary outcomes in ophthalmic trials.

Figure 2: Early Treatment of Diabetic Retinopathy Study (ETDRS) Chart
**AI Enhanced Image Based Upon Image Provided by Anna Tichenor, OD, PhD*

Eye Surface Damage

Corneal staining is a measure of eye surface health and damage, and it is one of the most used outcomes in the dry eye clinical trials space.[15–17] Corneal staining is performed by first applying an ocular surface dye to the eye – typically sodium fluorescein.[18,19] Historically, sterile fluorescein strips are wet with sterile saline, shaken to remove excess solution and applied to the eye's surface by lightly touching the strip to the eye's surface in a manner that avoids induced staining.[18] The eye is then evaluated with a lighted microscope to see if there is corneal staining present.[18] While informative, this approach makes judging disease severity highly subjective; therefore, this test has been standardised in several ways. Liquid fluorescein is now often used and applied with a volumetric pipette to control for fluorescein applied to the eye.[18] A special filter and cobalt blue light is now used when viewing the eye with the lighted microscope to help standardise and help visualise the corneal staining.[18] A standard amount of time (e.g., 1 minute) after applying the fluorescein is used for determining when to evaluate the cornea because the fluorescein itself dissipates on the eye with time.[18] And possibility the greatest advancement that has standardise this method is the use of grading scales. Lemp introduced one of the first corneal staining grading scale in the 1995 as part of the National Eye Institute's Industry Workshop on Dry Eye.[18] The original National Eye Institute corneal grading scale uses a 0 to 3 integer grading scale for each eye region, and it has four gestalt drawings to illustrate the amount of staining on the cornea with 0 representing no staining and 3 representing severe staining. The investigator then evaluates 5 corneal regions that have been pre-defined by the scale and a total score is obtained by adding the 5 regions together (0 to 15 scale with 15 being the worst possible grade) [18] This scale has been further enhanced by switching from gestalt grading to counting the number of staining dots and morphology of the cornea to help avoid investigator biases and to make the grades across doctors more repeatable.[20] These newer grading scales likewise have made it easier for Investigators to identify disease severity differences by making smaller grading increments (0.5-unit vs. 1.0-unit grading steps), which are still clinically meaningfully different and by expanding the grading scale (0–20 scale vs. 0–15 scale).[20]

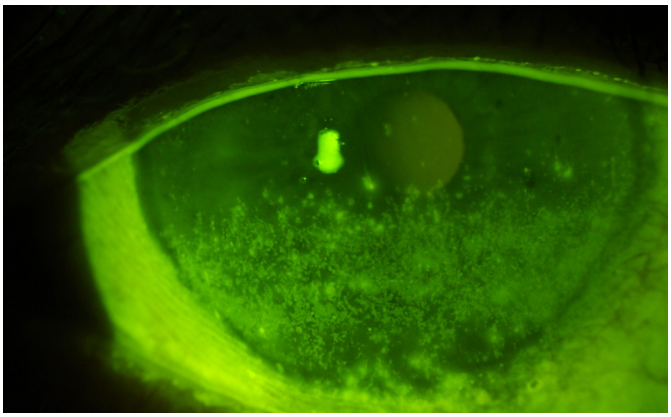

Figure 3: Example Image of an Eye Being Evaluated with Sodium Fluorescein
**Image Provided by Anna Tichenor, OD, PhD*

Endpoints Education

Educating the trial team on how to complete an endpoint and educating a participant on how to complete a test is essential because poor protocol compliance can result in endpoint variability.[2] With educating the trial team, several approaches can be taken such as having the team first train on the protocol. If an endpoint is essential for trial success, adding a knowledge check such as a multiple-choice test can help with understanding competency and if retraining is needed.[2] When an endpoint method is complex, simulations and in person practical tests can further help enforce trial standards and determining if the trial team fully understands the measure.[3] When it comes to the participant, it is also important to fully inform them about the test that are about to complete, so they are prepared during the testing situation. Procedure coaching can be provided assuming it is not going to bias data collection (e.g., avoid suggesting answers for questionnaires).

Key Factors Statistically Influencing Endpoint Success

We should care deeply about standardising trial endpoints because it is not only good scientific practice, but it will statistically result in smaller sample sizes (e.g., cheaper, faster trials) and because it will make it more likely that a trial will meet its endpoints. This should

start with selecting endpoints and optimising the endpoints to have good inter-examiner repeatability, which is a measure of how similarly two examiners makes the same measurement and intra-examiner repeatability, which describes how similarly one examiner makes the same measurement on two different occasions.[21] Accomplishing these goals subsequently results in better/less measurement variability, which helps statistically as highlighted by the following examples.

Let's consider theoretical corneal staining data from two different Phase II dry eye trials. Participants in each trial are seen at a baseline visit and a follow up visit and treated with a novel drug starting at the end of the baseline visit. Their right eyes are evaluated with 0 to 20 corneal staining scale at each visit. Table 1 provides example data from two different studies. Trial 1 has exceptionally low variability across participants while Trial 2 has higher variability across participants.

Participant	Trial 1		Trial 2	
	Visit 1	Visit 2	Visit 1	Visit 2
1	10.1	9.5	8.0	7.0
2	10.2	9.4	5.0	15.0
3	10.1	9.5	12.0	11.0
4	10.2	9.4	13.0	7.0
5	10.1	9.5	7.0	9.0
6	10.2	9.4	11.0	10.0
7	10.1	9.5	12.0	4.0
8	10.2	9.4	13.0	12.0
9	10.1	9.5	4.0	9.0
10	10.2	9.4	11.0	14.0
11	10.1	9.5	12.0	7.0
12	10.2	9.4	3.0	12.0
13	10.1	9.5	10.0	9.0
14	10.2	9.4	11.0	1.0
15	10.1	9.5	12.0	11.0
16	10.2	9.4	13.0	12.0
17	10.1	9.5	10.0	5.0
18	10.2	9.4	11.0	10.0
19	10.1	9.5	12.0	12.0
20	10.2	9.4	13.0	12.0
	Sum	Sum	Sum	Sum
	203	189	203	189
	Mean ± Standard Deviation		Mean ± Standard Deviation	
	10.15 ± 0.05	9.45 ± 0.05	10.15 ± 3.10	9.45 ± 3.47
	P-Value		P-Value	
	<0.0001		0.53	

Table 1: Theoretical Right Eye Total Corneal Staining Scores for 2 Studies

Visits 1 and 2 in for each trial have the same sum total and mean corneal staining score, yet the standard deviation (measure of variability) for the visits in Trial 1 are much lower than the standard deviation for the visits in Trial 2. If you then compare the two groups in Trial 1 and compare the two groups in Trial 2 with a statistical test (paired t-test), the post-treatment difference in Trial 1 is highly significantly different compared to baseline, but it is not significant in Trial 2.

If you now use these data to plan a follow up trial, you will use the standard deviations from your results in your sample size calculation and you would also need to determine what you would consider to be a clinically meaningful difference between groups (can the investigator notice this difference in the clinic). If we would again, plan a trial that is aimed to evaluate post-treatment change within the same participant, you would get the theoretical sample sizes described in Table 2.

Meaningful Difference	Trial 1	Trial 2
	Sample Size	Sample Size
0.5	3	304
1.0	2	78
1.5	2	36
2.0	2	21

Table 2 highlights that increased variability can have a major impact on sample size. It also highlights the importance granularity of the scales we use. With this example, if we use a 0.5-unit increment scale, which is included in some corneal staining scales, it has the potential to shrink samples size compared to using a scale with 1.0-unit increments.

Conclusion

Defined and highly standardised trial endpoints, when paired with robust methodological training, can play a critical role in improving trial success. These concepts will be critical when optimising existing endpoints and with developing new endpoints, which is of great current interest in the community. By ensuring that trial teams have a consistent understanding of testing protocols, variability in clinical trial data can be significantly reduced. This increased consistency enhances data quality, improves the reliability of trial results, and ultimately increases the likelihood of clinical trial success.

REFERENCES

1. Takebe T, Imai R, Ono S. The Current Status of Drug Discovery and Development as Originated in United States Academia: The Influence of Industrial and Academic Collaboration on Drug Discovery and Development. Clin Transl Sci. 2018;11(6):597-606.
2. Kobak KA, Engelhardt N, Williams JB, Lipsitz JD. Rater training in multicenter clinical trials: issues and recommendations. J Clin Psychopharmacol. 2004;24(2):113-7.
3. Huang P, Miao W, Wang R, Yang F, Li X, Shen N. Innovative multimodal educational strategies: assessing the impact of integrative teaching methods on standardized neurology resident training. BMC Med Educ. 2025;25(1):1032.
4. Wolffsohn JS, Benitez-Del-Castillo JM, Loya-Garcia D, Inomata T, Iyer G, Liang L, et al. TFOS DEWS III: Diagnostic Methodology. Am J Ophthalmol. 2025;279:387-450.
5. Nichols KK, Begley CG, Caffery B, Jones LA. Symptoms of ocular irritation in patients diagnosed with dry eye. Optom Vis Sci. 1999;76(12):838-44.
6. Maxwell C. Sensitivity and accuracy of the visual analogue scale: a psycho-physical classroom experiment. Br J Clin Pharmacol. 1978;6(1):15-24.
7. Carlsson AM. Assessment of chronic pain. I. Aspects of the reliability and validity of the visual analogue scale. Pain. 1983;16(1):87-101.
8. McMonnies CW. Measurement of Symptoms Pre- and Post-treatment of Dry Eye Syndromes. Optom Vis Sci. 2016;93(11):1431-7.
9. Schaumberg DA, Gulati A, Mathers WD, Clinch T, Lemp MA, Nelson JD, et al. Development and validation of a short global dry eye symptom index. Ocul Surf. 2007;5(1):50-7.
10. Azzam D, Ronquillo Y. Snellen Chart. StatPearls. Treasure Island (FL)2026.
11. Kaiser PK. Prospective evaluation of visual acuity assessment: a comparison of snellen versus ETDRS charts in clinical practice (An AOS Thesis). Trans Am Ophthalmol Soc. 2009;107:311-24.
12. Shamir RR, Friedman Y, Joskowicz L, Mimouni M, Blumenthal EZ. Comparison of Snellen and Early Treatment Diabetic Retinopathy Study charts using a computer simulation. Int J Ophthalmol. 2016;9(1):119-23.
13. Lim LA, Frost NA, Powell RJ, Hewson P. Comparison of the ETDRS logMAR, 'compact reduced logMar' and Snellen charts in routine clinical practice. Eye (Lond). 2010;24(4):673-7.
14. Early Treatment Diabetic Retinopathy Study (ETDRS): Manual of Operations. U.S. Dept. of Commerce, National Technical Information Service: 1985.

15. Terry RL, Schnider CM, Holden BA, Cornish R, Grant T, Sweeney D, et al. CCLRU standards for success of daily and extended wear contact lenses. Optom Vis Sci. 1993;70(3):234-43.
16. Joyce PD. Corneal vital staining. Ir J Med Sci. 1967;6(500):359-67.
17. Dundas M, Walker A, Woods RL. Clinical grading of corneal staining of non-contact lens wearers. Ophthalmic Physiol Opt. 2001;21(1):30-5.
18. Lemp MA. Report of the National Eye Institute/Industry workshop on Clinical Trials in Dry Eyes. CLAO J. 1995;21(4):221-32.
19. Korb DR, Greiner JV, Herman J. Comparison of fluorescein break-up time measurement reproducibility using standard fluorescein strips versus the Dry Eye Test (DET) method. Cornea. 2001;20(8):811-5.
20. Sall K, Foulks GN, Pucker AD, Ice KL, Zink RC, Magrath G. Validation of a Modified National Eye Institute Grading Scale for Corneal Fluorescein Staining. Clin Ophthalmol. 2023;17:757-67.
21. Powell DR, Nichols JJ, Nichols KK. Inter-examiner reliability in meibomian gland dysfunction assessment. Invest Ophthalmol Vis Sci. 2012;53(6):3120-5.

Andrew D. Pucker

Dr. Andrew Pucker is the Chief Development Office of Mintra Health. He is a world-renowned clinician-scientist with deep expertise in ophthalmic disease and clinical trial design. He earned his OD, MS and PhD degrees from The Ohio State University, and he is an adjunct professor at the University of Alabama at Birmingham. Dr. Pucker's independent research career has focused on ocular surface disease, contact lenses and myopia development.

Email: andrew.pucker@mintra.bio