Published in final edited form as : *Eur J Immunol.* 2016 Oct;46(10):2481-2483. doi: 10.1002/eji.201546272

Refined clinical scoring in comparative EAE studies does not enhance the chance to observe statistically significant differences

Silvia M. Tietz[1], Marcel Zwahlen[2], Neda Haghayegh Jahromi[1], Pascale Baden[1], Ivana Lazarevic[1], Gaby Enzmann[1] and Britta Engelhardt[1]

[1]Theodor Kocher Institute and [2]Institute of Social and Preventive Medicine, University of Bern, 3012 Bern, Switzerland

Key words: experimental autoimmune encephalomyelitis, clinical scoring, area under the curve, statistical analysis, Bland-Altman plots

Abbreviations:

aEAE: actively induced EAE (aEAE); **AUC**: area under the curve; **EAE**: experimental autoimmune encephalomyelitis; **MS**: multiple sclerosis; **tEAE**: adoptive transfer EAE; **3R**: Three Rs (reduction, refinement and replacement of antimal experimentation)

Experimental autoimmune encephalomyelitis (EAE) in rodents is frequently used as an animal model for multiple sclerosis (MS). A potential role of individual genes in MS pathogenesis is regularly investigated by comparing EAE development in gene-targeted mice with that in their wild-type littermates [1]. Clinical disease in EAE models ultimately amounts to motor dysfunction, thus EAE severity is generally scored in correlation to the severity of paralysis. However, there is no international standard scoring system, which research groups

would use to measure EAE severity and clinical assessment scales mostly in use reach from 0-3 points to 0-6 points [2-5]. The use of different EAE scoring systems prohibits direct comparison of clinical EAE data published from different laboratories. Different scoring might include differences in the subjective bias influencing the scoring and might also lead to differences in the results reported from the different laboratories, including differences in the statistical significance reported. Refined clinical EAE scoring includes sophisticated motor skill testing, which increases mouse handling time causing increased distress for the animals. Considering the 3R rules, aiming to refine animal experiments, we here asked if extended scoring of EAE in mice will grant improved documentation of clinical EAE allowing to perform a more powerful statistical analysis and thus identify more subtle differences when comparing EAE courses in different groups of mice. To this end we compared EAE scores in wild-type SJL/J and C57BL/6J mice with different *knockout* littermates, respectively using a 0-3 point scale used in our laboratory [6, 7] (*Materials and Methods* in Supporting Information) and a second, more refined scoring system [8, 9], which includes sophisticated evaluation of motor skill testing such as grid walk, a righting test (the ability of mice to turn over when they have been placed on their back) and a hanging test to determine front limb weakness. The latter score also included unilateral disease signs, which were scored as intermediate steps of 0.5 (*Material and Methods*). Over a time period of 9 months, 6 different researchers evaluated the disease progression in actively induced EAE (aEAE) models in both, C57BL/6J and SJL/J mice as described before [7] as well as in an adoptive transfer EAE (tEAE) model induced by the transfer of 2D2 T cells into C57BL/6 mice [10] in comparison to their respective *knockout* littermates using the two scoring systems in parallel. During the time of observation a total of 10 different mouse strains (C57BL/6 and SJL/J 2 wild-type mouse lines and 8 *knockout* mouse lines) were analyzed in 3 independent tEAE (n = 46) and 5

independent aEAE experiments (n = 157) and were included into the study. Graphical display of the disease course of not significantly different aEAE experiments (Fig. 1A), significantly different aEAE experiments (Fig. 1 E), and the tEAE experiments (Fig. 1I) evaluated with a 3-point scale showed a slightly less soft curve when compared to the curve revealed by the 5.5-point scale (Fig. 1C, G, K). The area under the curve (AUC) of the overall disease severity (Fig. 1B, D, F, H, J, L) was calculated for each mouse for the 3-point and the 5.5-point scale to compare the disease course of wild-type and knockout mice using non-parametric statistical tests (Mann Whitney test, also referred to as Wilcoxon rank-sum test). Despite a clear visual difference of the graphs with the 5.5-point scale and the graph with the 3-point scale, the obtained p-values were hardly different or even identical. Thus, we next asked whether the two scoring systems are really providing fundamentally different information. To address this we compared the discrepancies of the two scoring measurement methods by constructing Bland-Altman plots [11]. In these plots, the means of the two measurement systems on the same observation (on x-axis) are plotted against the differences of the two measurements (on y-axis). For these plots we rescaled the scorings from the 5.5-point scale to have a maximum of 3 (as the 3-point scale scorings) and similarly the corresponding area under the curve values (Figure 2 A-C). Additionally we constructed Bland-Altman plots for the relative ranks of the AUC derived from the 3-point and 5.5-point scorings (Fig. 2 D-F) for all three experiments. We found that there was a high agreement between the two measurement methods, suggesting that the amount of information of the two scoring systems is comparable, especially the relative rankings were very similar which then, in turn, resulted in very similar p-values when comparing wild-type with knockout mice with non-parametric statistical tests. Finally, statistical analysis of additional parameters e.g. mean day of onset, mean maximal disease score and mean day of resolution using the two different

scorings also produced similar or identical p-values (Supporting Information Fig. 1). Taken together, our results show that although a refined EAE scoring improves graphic representation of different clinical EAE scores between two groups, it failed to relevantly improve the ability to statistically establish differences between the overall disease course of the two groups of mice. Based on these observations we propose that 3-point scoring of disease progression in EAE is preferable to refined scoring systems as it comes, for almost the same amount of information, with less distress for the assessed mice and is thus more in line with the 3R guidelines.

References

1. Robinson, A.P., et al., *The experimental autoimmune encephalomyelitis (EAE) model of MS: utility for understanding disease pathophysiology and treatment.* Handb Clin Neurol, 2014. **122**: p. 173-89.
2. Mendel, I., N. Kerlero de Rosbo, and A. Ben-Nun, *A myelin oligodendrocyte glycoprotein peptide induces typical chronic experimental autoimmune encephalomyelitis in H-2b mice: fine specificity and T cell receptor V beta expression of encephalitogenic T cells.* Eur J Immunol, 1995. **25**(7): p. 1951-9.
3. Sobel, R.A., et al., *Acute experimental allergic encephalomyelitis in SJL/J mice induced by a synthetic peptide of myelin proteolipid protein.* J Neuropathol Exp Neurol, 1990. **49**(5): p. 468-79.
4. Westarp, M.E., et al., *T lymphocyte line-mediated experimental allergic encephalomyelitis--a pharmacologic model for testing of immunosuppressive agents for the treatment of autoimmune central nervous system disease.* J Pharmacol Exp Ther, 1987. **242**(2): p. 614-20.
5. Miller, S.D. and W.J. Karpus, *Experimental autoimmune encephalomyelitis in the mouse.* Curr Protoc Immunol, 2007. **Chapter 15**: p. Unit 15 1.
6. Abadier, M., et al., *Cell surface levels of endothelial ICAM-1 influence the transcellular or paracellular T-cell diapedesis across the blood-brain barrier.* Eur J Immunol, 2015. **45**(4): p. 1043-58.
7. Doring, A., et al., *E- and P-selectin are not required for the development of experimental autoimmune encephalomyelitis in C57BL/6 and SJL mice.* J Immunol, 2007. **179**(12): p. 8470-9.
8. Bischof, F., et al., *Specific treatment of autoimmunity with recombinant invariant chains in which CLIP is replaced by self-epitopes.* Proc Natl Acad Sci U S A, 2001. **98**(21): p. 12168-73.
9. Tietz, S.M., et al., *MK2 and Fas receptor contribute to the severity of CNS demyelination.* PLoS One, 2014. **9**(6): p. e100363.
10. Krishnamoorthy, G., et al., *Myelin-specific T cells also recognize neuronal autoantigen in a transgenic mouse model of multiple sclerosis.* Nat Med, 2009. **15**(6): p. 626-32.
11. Bland, J.M. and D.G. Altman, *Statistical methods for assessing agreement between two methods of clinical measurement.* Lancet, 1986. **1**(8476): p. 307-10.

Published in final edited form as : *Eur J Immunol.* 2016 Oct;46(10):2481-2483. doi: 10.1002/eji.201546272

Acknowledgemets

This work was funded by the Swiss National Science Foundation, the Swiss Multiple Sclerosis Society and the EU FP7 ITN nEUROinflammation to BE and an UniBe Initiator Grant to ST.

Conflict of interest

The authors declare no financial or commercial conflict of interest,

Figure legends

Figure 1. Comparison of two different EAE scoring protocols to assess the severity of EAE.

(A-H) Active EAE (aEAE) was induced in (A-D) wild-type SJL/J mice and *knockout* (KO) littermates and (E-H) in wild-type C57BL/6 mice and KO littermates as described [7]. (I-K) Adoptive transfer EAE (tEAE) using 2D2 TCR transgenic $CD4^+$ T cells was induced as described [10]. Representative disease courses of both aEAE and tEAE are shown. (A, C) The disease course of EAE was evaluated with (A) a 3-point scale and (C) a 5.5-point scale as well as (E, G) the disease course of a significantly different aEAE was evaluated with (E) a 3-point scale and (G) a 5.5-point scale, and (I, K) the disease course the tEAE was evaluated with (I) a 3-point scale and a (K) 5.5-point scale, and. (B, D, F, H, J, L) The severity of the disease was analyzed by the evaluation of the area under the curve (AUC). (A-D) Fifteen mice per group, data shown are from a single experiment representative of five independent experiments including a total of 157 mice. (E-H) Twelve C57BL/6 mice and six KO mice, data shown are from a single experiment representative of five independent experiments including a total of 157 mice. (I-K) Six mice per group, data shown are from a single experiment representative of three independent experiments including a total of 46 mice. (B, D, F, H, J, L) p values of the AUC were determined by Mann-Whitney test using GraphPad Prism 6 software. Data are shown as means ± SD.

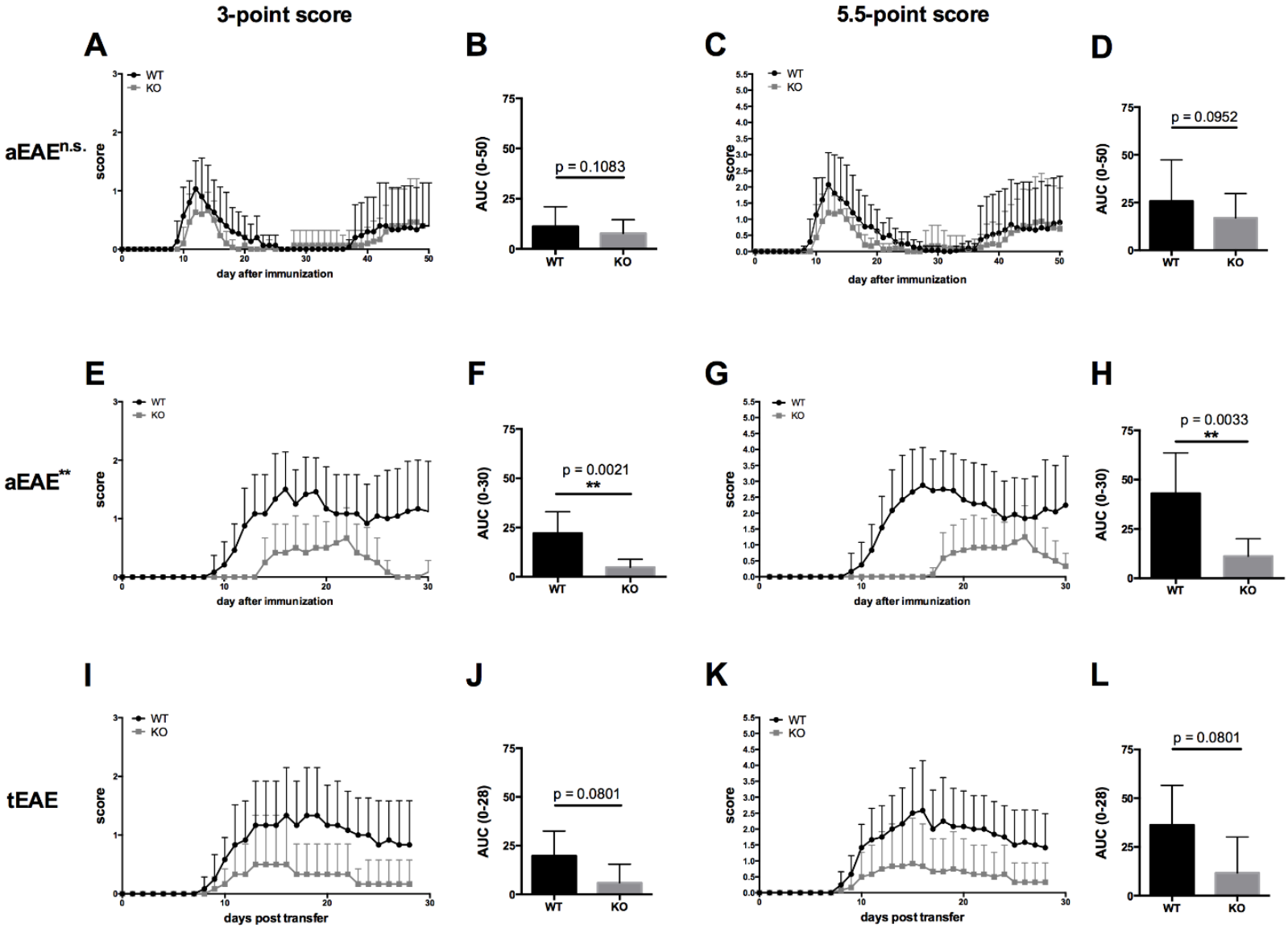
Figure 2. Bland-Altman plots of AUC values and of ranks of the AUC from the two scoring systems

Bland-Altman plots of the AUC derived from 3-point and rescaled 5.5-point scoring for active EAE induced in wild-type (WT) and knockout (KO) mice with (A) non-significant difference in disease severity and (B) significant difference in disease severity, and (C) with non-significant

difference in disease severity between tEAE in wild-type (WT) and knockout mice. (D to F) Bland-Altman plots for the ranks of the AUC derived from the 3-point and rescaled 5.5-point scoring for the same experiments as shown in (A to C) are shown. In Bland-Altman plots the means of the two measurement systems on the same observation (on x-axis) are plotted against the differences of the two measurements (on y-axis). The grey shaded area gives the 95% range for the difference between the two measurement systems.

Published in final edited form as : *Eur J Immunol.* 2016 Oct;46(10):2481-2483. doi: 10.1002/eji.201546272

Figure 1



Published in final edited form as : *Eur J Immunol.* 2016 Oct;46(10):2481-2483. doi: 10.1002/eji.201546272

Figure 2