High level appraisal of the Norwegian Institute of Public Health's report "Disease modifying treatments for relapsing remitting multiple sclerosis, including rituximab"

This document provides a brief overview of the key data and methodological considerations/limitations of the approach taken by the Norwegian Institute of Public Health (NIPH) in their clinical effectiveness/safety assessment in "Disease modifying treatments for relapsing remitting multiple sclerosis, including rituximab". It is not possible to know the details of all the methods used in the report due to lack of explanation and reporting. The assessment is based on our understanding of the report.

SUMMARY

The NIPH collected and reviewed the evidence for clinical effectiveness and general safety issues for disease modifying treatments for relapsing remitting multiple sclerosis, synthesised evidence from randomised controlled trials and non-randomised registry-based studies using network meta-analysis methods.

Upon review, a number of data and methodological limitations have been identified that require further consideration, these include:

- **General methods description** - reproducibility; the available information would not allow us to reproduce the analysis. Therefore, a thorough assessment of the methodology used is not possible. Such limited reporting does not support transparent decision making
- **Data** - publication bias and risk of bias assessment, lack of transparency on some data inputs to the comparative effectiveness;
    - Publication bias - non-randomised studies (NRS) are not typically registered with health authorities as randomised controlled trials are, therefore, may be subject to considerable publication bias.
    - Risk of bias - RCTs remain the gold standard in the hierarchy of evidence types to be used to address treatment efficacy, as randomisation balances unknown and unmeasured confounders. A high rating on a NRS risk of bias checklist (e.g. green, see Table 1 in NIPH report) does not change the fact that this is a NRS, and does not necessarily trump evidence to be gleaned from an RCT
    - Unpublished results were added to the comparative effectiveness modelling that lack supportive detail and were not subjected to a peer-review process. It is unclear how such data are to be assessed or evaluated as to their quality or appropriateness
- **Methods** - bias adjustment, downweighting and arm level (breaks randomisation); in some cases it is not clear how the NIPH addressed bias in the analysis and the NIPH's preferred analytical model (an arm-wise network meta-regression) breaks randomisation and is, therefore, discussed controversially in the literature.

Given the above, disease modifying therapies included in the NIPH assessment that rely largely on non-randomised data cannot be reliably included in evidence-based decisions.

General methods summary

Network meta-analysis (general high-level summary):
"Conventional meta-analysis synthesises evidence from studies that each compare a single pair of treatments (e.g., a treatment of interest versus placebo). NMA is a generalisation of conventional meta-analysis to the case where there are multiple treatments, and therefore multiple pairs of treatments that may be compared (91)." NIPH report p. 81.

Meta-regression (high-level summary):
Studies included in a meta-analysis may differ systematically in terms of populations, protocols, trial settings, etc. Meta-regression aims at adjusting for such differences and thereby reduce potential biases that may arise from these differences.
Meta-regression requires an understanding of the covariates expected to impact on treatment effects (so-called effect modifiers). Also, these covariates must be measured and reported in all studies synthesized in the meta-regression model. For a detailed discussion of meta-regression, see for example Chapter 8 in the book by Dias et al. (2018).[1]
NIPH used a meta-regression to model "systematic differences between randomized and non-randomized evidence" (p. 82 of NIPH report).

DETAILED EXPLANATION

- General methods description and reproducibility
- Publication bias
- Risk of bias assessment
- Lack of transparency on some data inputs
- Analysis of non-randomized studies
- Bias adjustment
- Downweighting
- Arm level (breaks randomisation)

General methods description and reproducibility

- The NIPH report provides a narrative summary of the analysis methods used. The primary network meta-regression model is outlined in the text, but no detailed description is given. In particular, a full (mathematical) model specification is lacking, and neither the analysis code nor the analysis data sets are given. The available information would not allow to reproduce the analysis. Therefore, a thorough assessment of the methodology used is not possible. Roche has asked for the full (mathematical) model specification, but this request was rejected by NIPH.

Publication bias

[1] Dias, Sofia, A. E. Ades, Nicky J. Welton, Jeroen P. Jansen, and Alex J. Sutton. 2018. *Network Meta-Analysis for Decision Making*. Wiley Series in Statistics in Practice. Hoboken, NJ: Wiley.

- Non-randomised studies (NRS) are not typically registered with health authorities as randomised controlled trials are (e.g. www.clinicaltrials.gov). As a result, relying on published NRS may mean that the evidence included in the network meta-analysis is biased towards positive and/or significant findings that the authors chose and managed to get into the public domain.[2]

Risk of bias assessment

- Although they can be subjective and lead to slightly different conclusions regarding the quality of a study (see Table 1 below), we do recognise the usefulness of checklist tools to assess the risk of bias in a given study. However, some checklists are designed for RCTs (e.g. Risk of bias tool from the Cochrane handbook[3] as used by NIPH report), and other checklists for NRS (e.g.checklist for cohort studies from the Handbook of Norwegian Institute of Public Health[4], as used in the NIPH report). A high rating on an NRS checklist (e.g. green, see Table 1 in NIPH report) does not change the fact that this is an NRS, and does not necessarily trump evidence to be gleaned from an RCT-even one ranked quite poorly on an RCT checklist itself (e.g. red, see table 1). RCTs remain the gold standard in the hierarchy of evidence types to be used to address treatment efficacy, as randomisation balances unknown and unmeasured confounders[5]. We find Table 1 to be misleading regarding the relative quality of different evidence types.

Lack of transparency on some data inputs

- Unpublished results were added to the comparative effectiveness modelling that lack supportive detail and were not subjected to a peer-review process. It is unclear how such data are to be assessed or evaluated as to their quality or appropriateness.

Analysis of NRS

- Randomization balances all confounding factors across treatment groups, whether observed or unobserved, known or unknown. Differences in outcomes between arms within a well conducted RCT are thus deemed to be the consequence of differences in treatments. Treatment comparisons from non-randomized studies come at an increased risk of bias since confounding factors may not be balanced across treatment groups, among other limitations. Sophisticated causal inference methods have been proposed in the literature (such as propensity scores or instrumental variables).[6] The NIPH report does

[2] Light RJ, Pillemer DB, *Summing up: the science of reviewing research*. Cambridge, MA: Harvard University Press, 1984; Rothman KR, Greenland S, *Modern Epidemiology Second* (2nd) Edition, Lippincott-Raven Publishers, Philadelphia PA, USA, 1998

[3] Chapter 8: Assessing risk of bias in included studies: Cochrane tools [cited December]. Available from: https://handbook-5-1.cochrane.org/chapter_8/8_assessing_risk_of_bias_in_included_studies.htm

[4] Folkehelseinstituttet. Slik oppsummerer vi forskning. Håndbok. 2014. Available from: https://www.fhi.no/kk/oppsummert-forskning-for-helsetjenesten/slik-oppsummerer-vi-forskning/

[5] McAlister F, Laupacis A, Wells GA, Sackett DL, Users' guides to the medical literature. XIX Applying clinical trial results. B Guidelines for determining whether a drug is exerting (more than) a class effect. *Journal of the American Medical Association* 1999;**282**: 1371-1377

[6] Rita Faria et al., "NICE DSU Technical Support Document 17:THE USE OF OBSERVATIONAL DATA TO INFORM ESTIMATES OF TREATMENT EFFECTIVENESS IN TECHNOLOGY APPRAISAL:

not mention such approaches, although they did identify a study that used propensity score matching. The reader cannot judge whether the NRS data have been analyzed appropriately to provide valid summary level inputs into the synthesis of RCT and NRS data in the meta-regression.

Bias-adjustment in NMA - were potential differences and systematic biases of NRS compared to RCTs adjusted for in the NMA?

- The outline of the meta-regression model describes an overall adjustment for potential systematic differences between NRS and RCT. It is not clear whether (in addition) specific imbalances in patient characteristics as measured in the available studies have been included. This seems particularly important given the arm-wise model preferred by the NIPH. While contrast based models (trial-level summaries) implicitly adjust for imbalances in prognostic factors between studies, arm-based models do not and differences between populations must be accounted for in the meta-regression model.
- Although it is unclear from the report, we cannot exclude the possibility that the model by NIPH is based on arm-based data from the trials which is then converted to a relative measure (or contrast) within the model which is then used in the comparison model, i.e. a contrast-based model with arm-based likelihood[7]. This approach is acceptable without adjustments for confounders if the model includes inputs from RCTs, but is not acceptable with NRS unless the arm-level data inputted into the model is indeed properly propensity-score adjusted.
- Meta-regression techniques can be appropriate for bias adjustment if external validity of a study is the primary concern (and if effect modifiers have been measured). In contrast, NRS often suffer from internal validity. In such cases, other bias adjustment techniques have been proposed.[8]
- The NIPH uses a random effect to account for heterogeneity between studies. Assumption in random effects models is that there are enough trials to estimate the between study heterogeneity[9]. The performance of random effect models is therefore affected when the number of studies is small[10]. This could be a particular issue for the estimated treatment effect of rituximab in the NMA, based on a single non-randomized study.

METHODS FOR COMPARATIVE INDIVIDUAL PATIENT DATA," accessed April 4, 2017, http://www.nicedsu.org.uk/Observational-data-TSD(2973296).htm.

[7] S. Dias and A. E. Ades, "Absolute or Relative Effects? Arm-Based Synthesis of Trial Data," *Research Synthesis Methods* 7, no. 1 (March 1, 2016): 23–28, https://doi.org/10.1002/jrsm.1184.

[8] Sofia Dias et al., *Network Meta-Analysis for Decision Making*, Wiley Series in Statistics in Practice (Hoboken, NJ: Wiley, 2018). (Chapter 9)
N. J. Welton et al., "Models for Potentially Biased Evidence in Meta-Analysis Using Empirically Based Priors," *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172, no. 1 (2009): 119–36, https://doi.org/10.1111/j.1467-985X.2008.00548.x.

[9] NICE DSU TECHNICAL SUPPORT DOCUMENT 18: http://nicedsu.org.uk/wp-content/uploads/2018/08/Population-adjustment-TSD-FINAL-ref-rerun.pdf

[10] The Stata Journal (2010) 10, Number 3, pp. 95–407
https://journals.sagepub.com/doi/pdf/10.1177/1536867X1001000307

Appropriate weighting of input data according to quality of evidence - were lower quality evidence sources down-weighted compared to high(er) quality inputs?

- Joint synthesis of a set of studies - whether RCTs or NRS - requires assumptions on how the different sources of evidence relate to the underlying "true" quantity we try to estimate. In this process, the quality of the different sources of evidence is judged with tools such as the GRADE system. In the analysis model, each data point should be weighted according to the quality grading attached to it. Failure to do so comes at a high risk of bias. For example, the GRADE definition of "very low certainty" states that "the true effect is likely to be substantially different from the estimate of effect", while "high certainty" stands for "we are very confident that the true effect lies close to that of the estimate of the effect". When synthesizing sources of evidence with different quality ratings (as judged from risk of bias and quality of included studies assessments), higher quality evidence should get larger weight. The NIPH report does not mention weighting of the input data according to evidence quality. The regression model did include terms for heterogeneity between randomized and non-randomized evidence, but it is not clear to what extent this approach achieved appropriate down-weighting of lower quality evidence.
- Karabis (2019) provides a succinct summary of methods for the joint synthesis of RCT and NRS data in network meta-analysis such as *design-adjusted analysis* and *multi-level hierarchical modeling*.[11] Such methods explicitly down-weight the NRS compared to RCT evidence.

Arm-based analysis model

- The NIPH analyzed most outcomes with three different network meta-analysis models (Table 11). The preferred model was an arm-wise network meta-regression. Two additional models (naive NMA and NMA of RCT evidence) were fitted as sensitivity analysis and used contrast-wise data. In the literature, meta-regression models are usually introduced as an extension to standard NMA and fitted to a similar data structure. This means also meta-regression models can be fitted to contrast-wise data. It is not clear why the NIPH preferred arm-level data for the meta-regression. Arm-level models "break randomization" and are therefore discussed controversially in the literature.[12]
- The NIPH report states that the arm-based network meta-regression "facilitates analysis of studies that form disconnected networks" (p. 82), without providing further details. Methods allowing to connect disconnected networks require strong assumptions, some of which are untestable, such as that all prognostic factors and effect modifiers are

[11] A. Karabis, "Network Meta-Analysis for Various Study Designs: Stepping Outside the Randomized Controlled Trials Comfort Zone Into the Real World," *Value & Outcomes Spotlight*, February 2019, https://www.ispor.org/publications/journals/value-outcomes-spotlight/abstract/january-february-2019/network-meta-analysis-for-various-study-designs-stepping-outside-the-randomized-controlled-trials-comfort-zone-into-the-real-world.
Susanne Schmitz, Roisin Adams, and Cathal Walsh, "Incorporating Data from Various Trial Designs into a Mixed Treatment Comparison Model," *Statistics in Medicine* 32, no. 17 (2013): 2935–49, https://doi.org/10.1002/sim.5764.
Orestis Efthimiou et al., "Combining Randomized and Non-Randomized Evidence in Network Meta-Analysis," *Statistics in Medicine*, January 1, 2017, n/a-n/a, https://doi.org/10.1002/sim.7223.

[12] S. Dias and A. E. Ades, "Absolute or Relative Effects? Arm-Based Synthesis of Trial Data," *Research Synthesis Methods* 7, no. 1 (March 1, 2016): 23–28, https://doi.org/10.1002/jrsm.1184.

accounted for in the model.[13] The NIPH report does not provide any further detail. Therefore, it is not possible to judge whether combination of disconnected networks is performed according to current best practices.

- There appears to be different uses of the term "arm-wise" model, and it is unclear which approach the NIPH refers to. In the second approach data from the trials is converted to a relative measure (or contrast) within the model which is then used in the comparison model. If such an analysis is used, it should not be necessary to do adjustments for confounders if only inputs from RCTs are used, but is not appropriate to do on NRS unless the arm-level data inputted into the model is properly propensity-score adjusted.

ROCHE RECOMMENDATIONS

Transparency and reproducibility

- We recommend to request further information from the NIPH to be able to reproduce the analysis (in particular the full model specification, input data sets for all outcomes, and analysis code).
- This would ensure full transparency regarding the methods used and allow the scientific community to assess the robustness of the NIPH findings. In particular, this would allow to assess the sensitivity of the results to model structure and assumptions.

Appropriate weighting of input data according to quality of evidence

- When synthesizing studies of different quality, we recommend to account for these quality differences by downweighting sources of lower quality compared to sources of higher quality evidence. Well established guidance on the hierarchy of evidence/levels of evidence comparing different study designs should be considered.[14] For example Karabis (2019) provides a succinct summary of suitable methods for the joint synthesis of RCT and NRS data in network meta-analysis.
- As a result, the uncertainty, as reflected in the width of confidence intervals, should be much smaller for comparisons relying on large amounts of high quality evidence compared to comparisons based on limited and/or low quality data.
- We expect this to result in large uncertainty in comparisons predominantly informed by low quality non-randomized evidence.

Arm-based analysis model

[13] D.M. Phillippo et al., “NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submission to NICE.,” *NICE DSU TSD18*, December 2016.
David M. Phillippo et al., “Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal,” *Medical Decision Making*, August 19, 2017, 0272989X17725740, https://doi.org/10.1177/0272989X17725740.
James E. Signorovitch et al., “Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research,” *Value in Health* 15, no. 6 (September 2012): 940–47, https://doi.org/10.1016/j.jval.2012.05.004.

[14] McAlister F, Laupacis A, Wells GA, Sackett DL, Users’ guides to the medical literature. XIX Applying clinical trial results. B Guidelines for determining whether a drug is exerting (more than) a class effect. *Journal of the American Medical Association* 1999;**282**: 1371-1377

- We recommend a further explanation on what sort of approach was taken in the arm-based analysis
- We recommend to perform meta-regression on contrasts, not on arm-level data, to ensure the synthesis relies on within study comparisons.
- The potential impact on the results cannot be judged based on the NIPH report itself. Full assessment would require running additional sensitivity analysis.

Connecting disconnected networks

- We recommend to connect disconnected networks via population matching methods described by Philippo et al.[15] instead of using arm-level meta-regression.
- The potential impact on the results cannot be judged based on the NIPH report itself. Full assessment would require running additional sensitivity analysis. However, connecting disconnected networks often comes at the cost of considerable uncertainty. Therefore, comparisons based on such methods should come with increased uncertainty.

[15] D.M. Phillippo et al., "NICE DSU Technical Support Document 18: Methods for Population-Adjusted Indirect Comparisons in Submission to NICE.," *NICE DSU TSD18*, December 2016.
David M. Phillippo et al., "Methods for Population-Adjusted Indirect Comparisons in Health Technology Appraisal," *Medical Decision Making*, August 19, 2017, 0272989X17725740, https://doi.org/10.1177/0272989X17725740.
James E. Signorovitch et al., "Matching-Adjusted Indirect Comparisons: A New Tool for Timely Comparative Effectiveness Research," *Value in Health* 15, no. 6 (September 2012): 940–47, https://doi.org/10.1016/j.jval.2012.05.004.

Table 1: Two different independent quality assessments of Spelman 2018[16] using two different NRS checklists and demonstrating subjectivity of answers to similar questions. (NB. Similar questions are listed side by side; though similar, the questions can differ slightly).

checklist used	**checklist for cohort studies from the Handbook of Norwegian Institute of Public Health [1] (as reported in NIPH 2019)**		**adaption of the Newcastle-Ottawa for Cross-Sectional studies, adapted by Herzog et al 2013 [2]**
Were the groups comparable for important background factors?	Yes	Is there sufficient description of the groups and distribution of prognostic factors?	Yes Patient baseline and disease specific characteristics are reported for both unmatched and matched cohorts, and were statistically analysed.
		Were the groups comparable on all important confounding factors?	No (prior to matching), Yes (after matching) Prior to matching, patients in the rituximab group were significantly older, had a higher baseline EDSS, longer disease duration, greater exposure to pre-baseline treatment and less relapse activity relative to unmatched interferon-beta/glatiramer acetate patients. Following propensity score matching, treatment groups were well balanced with regard to all baseline prognostic used to derive the propensity score: age, sex, EDSS, disease duration at baseline, pre-baseline DMT starts (number, number as a proportion of disease duration, index year), the proportion of disease duration on treatment, and relapse activity in the 12 and 24 months pre-baseline.
Were the exposed individuals representative of a defined population?	Yes		
Was the control group(s) selected from the same population as the exposed group(s)?	Yes		

[16] Spelman, T., Frisell, T., Piehl, F., & Hillert, J. (2018). Compar-ative effectiveness of rituximab relative to IFN-β or glatiramer acetate in relapsing-remitting MS from the Swedish MS registry. Multiple Sclerosis Journal, 24(8), 1087-1095

		Are the groups assembled at a similar point in their disease progression?	No prior to matching, Yes after matching Disease duration on study entry varied across unmatched groups, but was similar between treatment groups when patients were matched based on propensity scores derived using baseline characteristics.
Was the study prospective?	Yes		
Was exposure and outcome measured equally and reliably in the groups?	Yes	Is the intervention/ treatment reliably ascertained?	Unclear Treatment regimens were not described.
Were many enough people in the cohort followed-up?	Yes	What proportion of the cohort was followed-up?	Unclear Patient follow-up was not reported
An analysis of attrition was done to explain whether those who have abandoned the study differ from those who have been followed-up?	unclear	Were dropout rates and reasons for drop-out similar across intervention and unexposed groups?	No Overall discontinuation of treatment during the observation period was higher in interferon beta/ glatiramer acetate patients than in matched rituximab patients: 684 (74.2%) vs 37 (8%).
Was the follow-up time long enough to show positive and/or negative outcomes?	Yes		
		Was follow-up long enough for the outcomes to occur?	Unclear Mean on-treatment follow-up was reported, but the authors highlighted that the study would require a larger sample with longer patient-level follow-up to characterise confirmed disability progression.

Were known, possible confounding factors taken into account in the design and/or analysis of the study?	Yes	Was there adequate adjustment for the effects of these confounding variables?	Unclear Propensity score matching was used to balance baseline confounding, thus providing matched samples for analysis. However, a Rosenbaum sensitivity analysis was conducted for unmeasured confounding, which suggests that there may be confounding variables unaccounted for but it was confirmed that they were not likely to have changed any of the outcomes.
Was the person who assessed the results (endpoints) blinded to who was exposed and who was not exposed?	Yes	Was outcome assessment blind to exposure status?	Unclear Outcome assessment was not described.
		Was a dose-response relationship between the intervention and outcome demonstrated?	N/A

[1] Folkehelseinstituttet. Slik oppsummerer vi forskning. Håndbok. 2014. Available from: https://www.fhi.no/kk/oppsummert-forskning-for-helsetjenesten/slik-oppsummerer-vi-forskning/

[2] Herzog R, Alvarez-Pasquin MJ, Camino Diaz JLDB, Estrada JM, Gil A. Are healthcare workers' intentions to vaccinate related to their knowledge, beliefs and attitudes? a systematic review. BMC Public Health. 2013; 13(154): Available from: http://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-13-154