Current benefit assessment of ‚CARVYKTI' as an example of a need for AMNOG reform

The ‚Target Trial Emulation' for CAR-T in the AMNOG

The concept of ‚target trial emulation' and propensity score-based adjustment procedures are recommended methods for minimising systematic bias in indirect comparisons without a bridging comparator. Based on the comparison of the single-arm evidence for the CAR-T Ciltacel with data from the LocoMMotion study, which was recognised in the medical community as the best prospective study on the standard of care for multiple myeloma with multiple prior treatments, the application of the concept in the AMNOG is presented. The comparison and the dramatic survival benefit shown with Ciltacel, the validity and clinical significance of which is confirmed by medical experts, contrasts with the complete rejection by the G-BA. This discrepancy is worth looking back on.

>> The chimeric antigen receptor (CAR) cell therapy Ciltacabtagen Autoleucel (short: Ciltacel, trade name: ‚CARVYKTI') from the manufacturer Janssen-Cilag GmbH, an orphan drug, is authorised for the treatment of relapsed and refractory multiple myeloma after at least three prior therapies. The authorisation is based on the single-arm phase 1b/2 study CARTITUDE-1 (CAR1). The prospective non-interventional clinical trial LocoMMotion (LOCO) with the same inclusion criteria was initiated for a valid and meaningful comparison with the current standard of care using the target trial emulation concept. In addition, the first data from a randomised controlled trial (RCT) on Ciltacel are available with CARTITUDE-4 (CAR4). Part of the study population falls within the currently authorised area of application and was mergedwith the ‚Target Trial Emulation' to provide support .This evidence was the basis for the benefit assessment of the Federal Joint Committee (G-BA) with its decision in August 2023. In the following, the analysis strategy is presented and the assessment by the G-BA is critically discussed based on the outcome overall survival.

The ‚target trial emulation' of the CAR1 and LOCO studies

The open, multicentre, single-arm study CAR1 comprises N=126 patients1, of which N=124 patients fulfil the criteria of the approved indication. Data with a median follow-up duration of 28.55 months were available in the benefit assessment.

The non-interventional clinical trial LOCO comprises N=248 patients, of whom N=110 fulfil both the criteria of the approved indication and the threshold values for laboratory parameters considered in CAR1 and receive a therapy defined by the G-BA as comparator therapy. Data were available with a median follow-up duration of 26.9 months.

The RCT CAR4 compares Ciltacel with the combination therapies daratumumab-pomalidomide-dexamethasone or pomalidomide-bortezomib-dexamethasone of the physician's choice and comprises N = 208 vs. N = 211 lenalidomide-refractory patients with one to three prior therapies, of whom N = 20 vs. N = 29 fulfil the criterion (at least three prior therapies) of the approved indication. Data from the first interim analysis with a median follow-up duration of 15.9 months were available. In principle, the pooling of evidence from randomised and non-randomised studies is advocated to increase the informative value (2). The pooled analyses thus provide an opportunity to assess the consistency of the resulting effects.

For the ‚target trial emulation' from CAR1 and LOCO, this results in an analysis population of N = 124 vs. N = 110 patients. With additional consideration of CAR4, N = 144 vs. N = 139 patients are in the analysis population.

Statistical methods for adjusting the comparison groups

The unadjusted and various adjusted analyses are presented. The adjustment for confounders was performed using propensity score (PS)-based methods, which are established as standard procedures in the scientific literature (3). A PS value was calculated for each study participant based on the individual characteristics of the confounders considered. The resulting distribution of PS for the treatment and control groups and the area of overlap between these two PS distributions provide an indication of the comparability of the groups. The various methods used address different aspects of potential bias, so that the robustness and significance of the results can be assessed in the overall view of all analyses.

Evaluation of the quality and significance of target trial emulation

In order to assess the validity of the comparison of Ciltacel with the standard of care in the context of target trial emulation , the quality of the evidence and, in particular, the positivity resulting from the comparability of the study populations must be examined and discussed in addition to the clinical relevance.

Positivity exists if each patient in the analysis population has a positive probability of being assigned to both the intervention group and the control group according to the expression of their characteristics. Further diagnostic criteria for evaluating the methodological quality of the comparison are overlap of the PS distribution and balance

---

1: Includes all apherised patients from both cohorts (USA and Japan)

„Target Trial Emulation"

Target trial emulation involves replicating an RCT on the basis of non-randomised data. The concept was developed by Hernán and Robins in 2016 to minimise the limitations of non-randomised, indirect comparisons (1). Since then, the concept has also been recommended by IQWiG for benefit assessments, e.g. using healthcare-related data (2).

An adequate target trial design, taking into account various characteristics of randomised trials (Table 1), can maximise the structural similarity between the populations to be compared, which is crucial for the validity of a comparison. While structural equality is guaranteed in randomised comparisons apart from random fluctuations, in indirect comparisons the extent of structural equality must be checked and, if necessary, increased by methodological or statistical measures. In the context of ‚target trial emulation' , a comprehensive evaluation and adjustment of structural equality is possible, e.g. with the help of PS-based methods.

Zusammenfassung der Komponenten für die „Target Trial Emulation“	
Protokollkomponente	**Beschreibung**
Einschlusskriterien	Die Einschlusskriterien der Studien sind weitestgehend identisch. Aus der Studie LOCO wurden Patienten ausgeschlossen, die nicht die in der Studie CAR1 festgelegten Schwellenwerte für Laborparameter erfüllten.
Zu untersuchende Behandlung	Die Patienten erhielten Ciltacel zulassungskonform. Aus der Studie LOCO wurden Patienten ausgeschlossen, die keine der vom G-BA definierten adäquaten Vergleichstherapien erhalten haben.
Behandlungszuweisung	Um die für einen Vergleich erforderliche Strukturgleichheit annähern zu können, muss eine Adjustierung relevanter Confounder erfolgen (3). Die Identifikation von Confoundern erfolgte in mehreren Schritten mithilfe systematischer Literaturrecherche und Experteninterviews und wird ausführlich im Dossier zur Nutzenbewertung beschrieben. Relevante Confounder (s. Tab. 4-37 im Nutzendossier) wurden in das Adjustierungsset für die vergleichende Analyse eingeschlossen (Hauptadjustierungsset). Im Rahmen einer im Dossier dargestellten Sensitivitätsanalyse wurden darüber hinaus alle identifizierten Confounder in das Adjustierungsset eingeschlossen (4).
Beobachtungsdauer	In der Studie CAR1 ist Baseline definiert als Zeitpunkt der Apherese. In der Studie LOCO ist Baseline definiert als Zeitpunkt der ersten Verabreichung der Studienmedikation. Die mediane Beobachtungsdauer ist vergleichbar.
Endpunkte	Die in den Studien CAR1 und LOCO erhobenen Endpunkte sind identisch operationalisiert.
Interessierender Vergleich	Das Analyse-Set entspricht dem ITT-Ansatz, da alle Studienteilnehmer gemäß Behandlungsgruppe analysiert wurden, unabhängig davon, ob sie diese Behandlung tatsächlich erhalten haben.
Analyseplan	Die vergleichende Analyse war in einem statistischen Analyseplan präspezifiziert und darauf ausgerichtet, den Anforderungen der Nutzenbewertung zu entsprechen.

Tab. 1: Summary of the components for target trial emulation from CAR1 and LOCO.

of patient characteristics. The overlap describes the proportion of patients from both treatment arms for whom PS of comparable size were calculated, taking into account the confounders. A deviation from the equal distribution of PS is not to be equated with a lack of positivity. Balancedness expresses the structural equality of the comparison groups with regard to the expression of the confounders after application of PS-based methods.

Before applying statistical methods, it must be checked that there is no structural violation of positivity. A structural violation exists if, for example, it is theoretically impossible to assign a treatment due to contraindications. A violation can also result from random fluctuations in the sample (5).

On the basis of expert knowledge, it must be critically examined to what extent the theoretical impossibility of treatment allocation (structural violation of positivity) can be ruled out, so that an adjustment is permissible against the background of structural conditions (6). The graphical representation of the PS distribution and the consideration of patient characteristics provide assistance in the assessment of positivity. Distributions with a small area of overlap as well as different manifestations of characteristics can be signs of a violation of the positivity assumption, which must be assessed by clinical experts. In addition, the reduction of bias due to confounding by the statistical models should be weighed against the influence of a possible violation of the positivity assumption (6).

Discussion of positivity and structural similarity in the Ciltacel comparison

The study protocol of the LOCO study is largely based on the CAR1 protocol. The inclusion criteria were largely adopted. One exception is threshold values for laboratory parameters, which were applied post hoc to the LOCO population. Therefore, structural positivity is given for the comparison between CAR1 and LOCO. The PS distributions of the naïve comparison show that patients from CAR1 and LOCO can be found in all PS intervals (Fig. 1A). Consequently, each patient in the analysis has a positive probability of receiving each of the two therapies being compared. The adjustment is therefore permissible. The description of the (pseudo) populations based on patient characteristics shows that a high degree of balance between CAR1 and LOCO could be achieved by adjustment (4). The PS distribution (Fig. 1) illustrates the increase in structural similarity through adjustment.

PS matching results emphasise the validity of the survival advantage

With the 1:1 PS matching method, pairs of patients from the control and intervention groups are formed based on the PS, which corresponds to the imitation of a 1:1 randomisation. Given the structural equality within the analysis populations to be compared in PS matching, the clear survival benefit of Ciltacel is confirmed (HR: 0.33, 95% CI: [0.18-0.62]). Against the background of the reduced number of cases in the matching (124 vs. 110 to 48 vs. 48), it must be evaluated to what extent the results of the PS matching population can be transferred to the overall population. There were no significant differences in the survival prognosis between the patients in the matched sample and the non-matched patients, neither for Ciltacel nor for the standard of care. The transferability is additionally underlined by the high consistency of the PS matching results, both in comparison to the naïve analysis and in comparison to the PS-weighted analyses, and thus allows reliable statements to be made about the existence of a survival advantage in favour of Ciltacel in the overall population.

Consistent analysis results show dramatic survival benefit with Ciltacel

The results for overall survivalpaint a homogeneous picture of a clear survival advantage in favour of Ciltacel across all adjustment methods and across both comparison constellations(Fig. 2). The very precise and consistent effect estimates across different analysis populations show that treatment with Ciltacel approximately triples the probability of survival compared to the standard of care, and are thus in the order of magnitude of a dramatic effect. In the context of non-randomised evidence, this underlines that the observed differences are actually due to the effect of the therapy studied and+ not to systematic bias (2). In addition, case number losses in matching and trimming do not lead to greater uncertainty, as the result is consistent across all adjustment methods and samples. The clear survival benefit with Ciltacel is also reflected in the relevant subpopulation of the RCT CAR4 (hazard ratio 0.42) despite the shorter follow-up period. Thus, indirect and direct evidence are consistent. A joint

Propensity Score-basierte Adjustierungsmethoden	
Average Treatment Effect (ATE)	Patienten des Kontrollarms (LOCO +/- CAR4) und des Interventionsarms (CAR1 +/- CAR4) werden basierend auf ihren PS mittels Inverse Probability Weighting (IPW) gewichtet, um Strukturgleichheit herzustellen. Hierbei wird eine Pseudo-Population generiert, welche die durchschnittlichen Charakteristika beider Vergleichsgruppen widerspiegelt.
Average Treatment Effect Among the Treated (ATT)	Nur Patienten des Kontrollarms (LOCO +/- CAR4) werden basierend auf ihren PS mittels IPW gewichtet, um Strukturgleichheit mit dem Interventionsarm herzustellen. Nach der Gewichtung wird auf die ursprüngliche Patientenzahl zurückskaliert. Hierbei wird eine Pseudo-Population generiert, welche die Charakteristika der Interventionsgruppe widerspiegelt.
Trunkierung und Trimmung bei ATE- und ATT-Analysen	Sehr große bzw. sehr kleine PS führen zu extremen Gewichten, die die Robustheit der statistischen Modelle beeinträchtigen können. Um extreme Gewichte zu adressieren, werden die Verfahren Trunkierung und Trimmung angewandt. Als Schwellenwerte werden hierbei die 2,5%- bzw. 97,5%-Perzentile der PS-Verteilung herangezogen. Bei der Trunkierung werden die Gewichte der Patienten, die außerhalb dieses Bereichs liegen, durch ein Minimal- bzw. Maximalgewicht entsprechend der 2,5%- bzw. 97,5%-Perzentile ersetzt. Bei der Trimmung werden Patienten, die jenseits der zuvor genannten Perzentile liegen, aus der Analysepopulation ausgeschlossen. Durch das Trimmen wird eine Pseudo-Population gebildet, in der Strukturgleichheit erhöht wird und stellt eine Annäherung an den Überlappungsbereich dar.
1:1-Matching	Beim PS-Matching werden Paare mit jeweils einem Patienten aus dem Ciltacel-Arm und einem Patienten aus dem Vergleichsarm gebildet, wobei jeweils ein Partner mit der optimalen Entfernung des PS gewählt wird. Die Paare bilden eine Pseudo-Population mit Strukturgleichheit, die einen Überlappungsbereich der PS-Verteilung darstellt.

Tab. 2: Overview of propensity score-based adjustment methods.

analysis with CAR4 was presented as part of the statement on the benefit assessment and also shows consistent results.

Clinical experts confirm comparability of CAR1 and LOCO

In the benefit assessment by the Department of Medicine, the G-BA assessed the LOCO study as unsuitable for an indirect comparison with the CAR1 study due to selection bias. One of the decisive factors for this was that, in addition to the criteria for the therapeutic indication, the suitability for CAR-T therapy must also be assessed individually before a decision is made in favour of CAR-T therapy. The G-BA sees this as an indication of a selection bias and, as a result, a lack of positivity. As further justification for the assumption of a selection bias, the G-BA cites the fact that the CAR1 and LOCO trials were conducted in different regions, and also refers to a recommendation of the S3 guideline to offer patients with multiple myeloma treatment within the framework of clinical trials. This may lead to a separation of patients who are suitable for either interventional clinical trials or application-oriented trials (4).

The question of whether a strong selection bias and consequently a lack of positivity can be assumed was the subject of the commenting procedure for the benefit assessment. In the oral hearing in particular, a clear confirmation of the comparability of the patient populations within the ‚target trial emulation' and the resulting validity of the comparison by the clinical experts emerged, in contrast to the G-BA's assessment. Furthermore, it emerged from the experts' statements that no causes of selection bias were seen in the conduct of the trials in different regions or in the recommendation to participate in clinical trials (4). Consequently, a structural violation of positivity cannot be assumed. This clinical view is decisive for the validity of the comparison (6, 7). The G-BA acknowledges the contrary assessment of the clinical experts, but nevertheless adheres to its original assumption of a lack of positivity in its final assessment (4).

Only part of the available evidence is used to assess the added benefit

As a result of the rejection, adjusted analyses are mentioned by the G-BA at best descriptively or used selectively. With regard to the criteria of overlap and balance, the submitted analyses are neither considered in their entirety nor weighed up by the G-BA. The evidence as a whole comprises naive and adjusted analyses in various cuts as well as additional pooled analyses with the CAR4 study. Almost perfect overlap and balance is achieved in particular through PS matching.

The distinction between structural and purely statistical violations of the positivity assumption described in the scientific literature is not made, although the G-BA cites this literature. The evidence was largely submitted in the commenting procedure at the request of the G-BA, but is not presented and discussed in a balanced manner. With regard to the analyses on PS matching and trimming, which are explicitly recommended in the literature in order to counter possible violations of the assumption of positivity and the resulting extreme weights, the G-BA only addresses the loss of case numbers as further alleged evidence of a lack of positivity, but not the statements on the transferability of the results of the pseudo-populations to the overall population. In particular, the G-BA does not address the consistency of all analyses. The consistent results from the direct evidence of CAR4 are also not included in the discussion. Limitations of the analysis, missing values and different data collection times of the adjustment variables are emphasised, but not placed in relation to the pronounced and consistent effects. The G-BA only refers to parts of the evidence and ignores important aspects of the overall view. By rejecting the evidence for the quantification of the added benefit, the G-BA negates, as it were, the existence of a treatment effect on overall survival at all. This is incomprehensible in view of the results presented and the course of the commenting procedure (4).

Effect size must be considered for the assessment of certainty of conclusions

In previous benefit assessments, the G-BA considered the magnitude of the effect size in order to assess the certainty of the indirect comparison with regard to potential systematic bias. In previous assessment practice, a sufficiently large effect in the sense of a dramatic effect, as shown by Ciltacel in overall survival, led to the inclusion of the indirect evidence in the decision-making process, even if no quantification of the added benefit was stated. This also applies in evidence situations in which neither an adequate confounder identification nor an adjustment of the indirect comparison was carried out, and was done without discussion of balance and positivity (8, 9). In the procedure on Ciltacel, on the other hand, the G-BA assumes such a strong systematic bias that the existence of a survival benefit through treatment with Ciltacel is fundamentally doubted.

The effect size in overall survival is not

considered for the assessment of certainty of conclusions in this procedure (4). <<

von: Janine Jakobs BS MS,
Ronja Kerßenboom BS MA,
Christina Schulat BA MS,
Dr. rer. nat. Jörn Sindern*
* Janssen Deutschland

Citation

Jakobs et al.: „The ‚Target Trial Emulation' for CAR-T in the AMNOG", in „Monitor Versorgungsfor-schung" (06/23), pp. 46–49. http://doi.org/10.24945/MVF.06.23.1866-0533.2563

PS distribution before adjustment (A) and after adjustment (B, C, D)

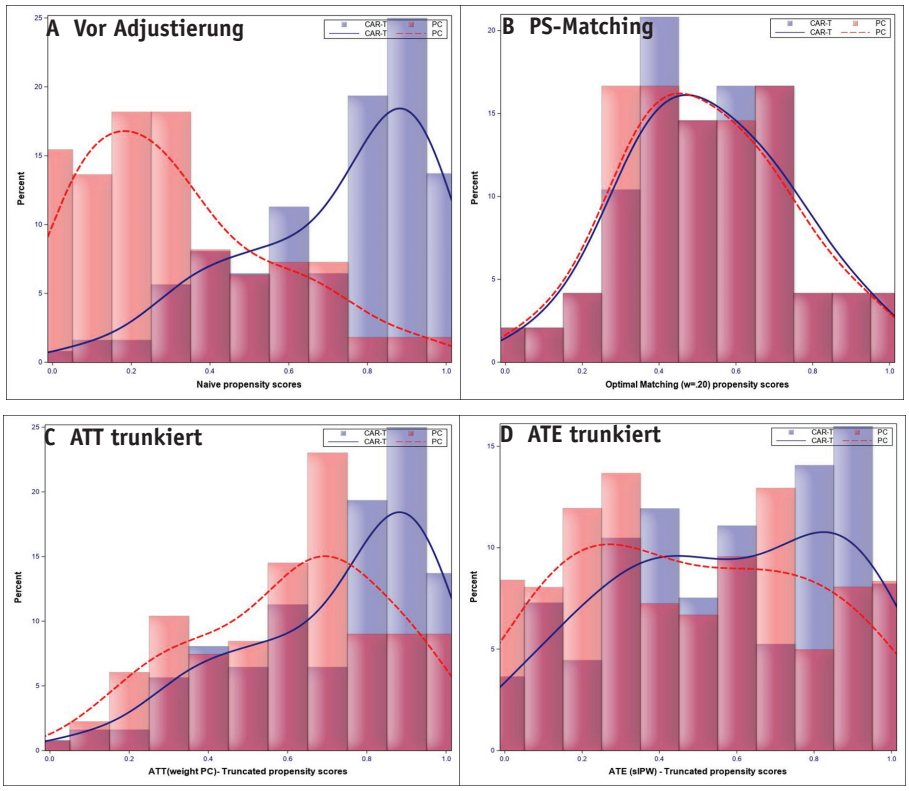


Fig. 1: The figure shows the PS distribution before adjustment (A) and after PS matching (B) and truncated ATT (C) or truncated ATE weighting (D). In the naïve analysis (A), there are differences between CAR1 (Ciltacel) and LOCO (PC; physician's choice), which are equalised by adjustment. PS matching in particular results in an analysis population with almost perfect overlap. Extreme weights in the ATT and ATE analyses were addressed by trimming and truncation.

Overall survival of Ciltacel versus standard of care

Analysis Performed	HR (95% CI)	P-value	Ciltacel N	PC N
Unadjusted	0.37 (0.25, 0.56)	<0.0001	124	110
IPTW - ATT	0.20 (0.13, 0.31)	<0.0001	124	110
IPTW - ATT (Truncated)	0.29 (0.18, 0.46)	<0.0001	124	53
IPTW - ATT (Trimmed)	0.31 (0.19, 0.50)	<0.0001	91	66
IPTW - ATE	0.29 (0.16, 0.53)	<0.0001	126	132
IPTW - ATE (Truncated)	0.31 (0.19, 0.49)	<0.0001	113	98
IPTW - ATE (Trimmed)	0.35 (0.21, 0.59)	<0.0001	89	58
PS Matching (c=0.20)	0.33 (0.18, 0.62)	0.0005	48	48

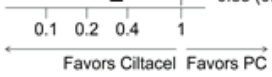


Fig. 2: Overall survival of Ciltacel versus the standard of care (PC; physician's choice) from CAR 1 and LOCO. The analysis of overall survival shows a high level of consistency across all adjustment methods. Weighted analyses were performed with variance correction.

Literature

1. Hernán, M. A., & Robins, J. M. (2016). Using big data to emulate a target trial when a randomized trial is not available. American journal of epidemiology, 183(8), 758-764.
2. IQWiG. Rapid Report. IQWiG-Berichte – Nr. 863. Version 1.1. Stand: 13.05.2020. 2020. https://www.iqwig.de/download/a19-43_versorgungsnahe-daten-zum-zweckeder-nutzenbewertung_rapid-report_v1-1.pdf
3. Hernán, M. A., & Robins, J. M. (2020). Causal Inference: What If; Chapman & Hall/CRC: Boca Raton, FL, USA, 85-100.Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. Statistical methods in medical research, 21(1), 31-54.
4. G-BA. Nutzenbewertungsverfahren zum Wirkstoff Ciltacabtagene autoleucel (Multiples Myelom, mind. 3 Vortherapien). https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/924/#beschluesse
5. Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., & Van Der Laan, M. J. (2012). Diagnosing and responding to violations in the positivity assumption. Statistical methods in medical research, 21(1), 31-54.
6. Cole, S. R., & Hernán, M. A. (2008). Constructing inverse probability weights for marginal structural models. American journal of epidemiology, 168(6), 656-664.
7. Zhu, Y., Hubbard, R. A., Chubak, J., Roy, J., & Mitra, N. (2021). Core concepts in pharmacoepidemiology: Violations of the positivity assumption in the causal analysis of observational data: Consequences and statistical approaches. Pharmacoepidemiology and drug safety, 30(11), 1471-1485.
8. G-BA. Tragende Gründe. Axicabtagen-Ciloleucel. 2. Mai 2019. https://www.g-ba.de/downloads/40-268-5741/2019-05-02_AM-RL-XII_Axicabtagen-Ciloleucel_D-406_D-416_TrG.pdf
9. G-BA. Tragende Gründe Idecabtagen vicleucel. 16. Juni 2022. https://www.g-ba.de/downloads/40-268-8579/2022-06-16_AM-RLXII_Idecabtagen%20vicleucel_D-779_TrG.pdf