PCR48

Common Criticism of Patient-Reported Outcomes (PRO) in Health Technology Assessments (HTA) by the Institute for Quality and Efficiency in Health Care (IQWiG) in Breast Cancer and Non-Small Cell Lung Cancer (NSCLC)

Jenya Antonova, MS, PhD[1], Jillian Lusk, MPH[2], Niclas Kürschner, MPH[3], Sarah Böhme, MS[3], Magdalena Harrington, PhD[2]
[1] Compass Strategy and Research, Inc. [2] Pfizer, Inc., [3] Pfizer Pharma GmbH

BACKGROUND

- Consistent with the Joint Clinical Assessments (JCA) requirements[1], German health technology assessment (HTA) is performed by Gemeinsamer Bundesausschuss (G-BA).
- Commissioned by G-BA, IQWiG evaluates the added benefit of pharmaceutical interventions based on the data reporting mortality, morbidity (effects of the disease and therapy), health-related quality of life (HRQoL), and adverse events, which are evaluated based on the certainty of the evidence base and the magnitude of effect[2]. All endpoints must be patient-relevant, and concepts like disease symptoms, HRQoL, and symptomatic adverse events should be evaluated via patient-reported outcomes (PROs).
- PRO data can determine[3] or improve,[4,5] the outcomes of benefit assessment in the context of acceptable adverse events profile. However, PRO data must be of good quality to be used in German HTAs.
- To elucidate IQWiG requirements for PRO data quality, we summarized IQWiG critique of PROs in assessments for metastatic HER2- breast cancer (H2NBC) and non-small cell lung cancer (NSCLC) submissions since year 2019.

METHODS

- From the IQWiG website (https://www.iqwig.de/en/projects/), we obtained and reviewed assessments, dated:
  - H2NBC: January 2019—August 2024,
  - NSCLC: January 2019—February 2025.
- For the assessments containing PRO data, we reviewed IQWiG feedback related to all data in the submission with in-depth focus on PROs.
- IQWiG critique of PRO data was summarized and categorized.

RESULTS

- The search identified the following assessments:
  - H2NBC assessments: among 17 assessments for 11 therapeutic agents, most occurred in 2019 (n=5) or 2020 (n=5), were for HER2- HR+ subtype (n=10; Figure 1), in the first (n=4) or first and second line (n=8; Figure 2).
  - NSCLC assessments: among 17 assessments for 12 therapeutic agents, most occurred in 2019 (n=6), were in the first line (n=15; Figure 3), and targeted epidermal growth factor receptor (EGFR-) and anaplastic lymphoma kinase (ALK-) negative NSCLC (n=9; Figure 4).
- Most assessments included several shortcomings.
- The shortcomings (by categories) and their consequences are detailed in Table 1 and summarized below:
  - Availability of data:
    - IQWiG shared negative comments when PRO data were absent from the randomized clinical trial analyses (n=14) or in adjusted indirect comparison (n=5).
  - Critique of clinical trial design:
    - Open-label trials resulted in a reduction regarding the certainty of the evidence base (n=16),
    - Critique of the observation period resulted in either reduction of quality of PRO data (n=10) or data rejection (n=4),
    - High levels of missing data resulted in a reduction in the certainty of the evidence base (n=9) or comments without negative consequences (n=1; all missing data were explained by death),
    - Discrepancies between study arms resulted in the reduction of PRO data quality (n=2) or data rejection (n=2).
  - Shortcomings of analyses:
    - Discrepancy of PRO analyses with the pre-defined plan resulted in data rejection (n=4),
    - Critique of data censoring resulted in the reduction of data quality (n=4) or data rejection (n=4),
    - Insufficient description of an endpoint resulted in the endpoint rejection (n=6),
    - Disagreement with an endpoint definition resulted in the rejection of that endpoint (n=19).
  - Relevance concerns:
    - In the absence of justification for selecting specific PRO-CTCAE items for assessment, PRO-CTCAE data were rejected (n=3),
    - In the absence of data on symptoms' seriousness and severity, the symptoms were categorized as non-serious and non-severe (n=11).
- In all instances, IQWiG rejected progression-free survival (PFS) data as not patient-relevant, citing the need to generate data demonstrating the long-term effect of the progression on PROs.

DISCUSSION

- Our review of IQWiG critique of PRO data revealed the critical importance the agency places on the quality of PRO data in the HTA submissions.
- Moreover, IQWiG applies rigorous requirements to data quality, reducing certainty of the evidence base or rejecting the data altogether when the quality standards were not met.
- To improve the utility of PRO data in IQWiG assessments, sponsors should consider the following strategies:
  - Enhance PRO data quality via appropriate clinical trial design (e.g., double-blinded design, balanced data collection between treatment arms, long-term PRO assessment post-progression),
  - Ensure high-quality PRO data collection in the trials (e.g., minimal missing data),
  - Generate evidence to support patient-relevance of endpoints,
  - Justify PRO-CTCAE symptoms selected for assessment,
  - Define symptom seriousness,
  - Adequately define endpoints in a pre-specified analysis plan and submit the plan along with the results.
- Given that the best practices require in-depth decisions regarding study design and data analyses, robust preparatory work is required prior to Phase 3 to ensure that relevant instruments are included in the pivotal trial and endpoints are adequately specified, powered, and analyzed. This preparatory work is recommended to be performed during early-phase clinical trials (e.g., Phase 1 and Phase 2).
- The limitation of this research include the following:
  - The current review was limited to H2NBC and NSCLC assessments between 2019 and August 2024 and February 2025 (see above). Therefore, some additional PRO related comments may have been included in assessments that fell outside of the current scope. Sponsors should conduct a careful assessment of IQWiG and G-BA requirements related to therapeutic area of interest to ensure adequate preparedness for upcoming HTA assessments.
  - IQWiG comments were not always detailed, making it difficult to interpret the intended request for best practices. Therefore, sponsors should consult with the HTA agencies (e.g., G-BA in Germany) directly before initiating pivotal trials to ensure that their strategy aligns with the expectations.

REFERENCES

1. Health Technology Assessment Coordination Group. Guidance on Outcomes for Joint Clinical Assessments. Published June 10, 2024. Accessed February 17, 2025. Available at: https://health.ec.europa.eu/document/download/a70a62c7-325c-401e-ba42-66174b656ab8_en?filename=hta_outcomes_jca_guidance_en.pdf.
2. Institute for Quality and Efficiency in Health Care (IQWiG). General Methods Version 7.0 of 19 September 2023. IQWiG; 2023. Available at: https://www.iqwig.de/en/general-methods-version-7-0
3. Gemeinsamer Bundesausschuss (G-BA). Benefit assessment procedure for the active ingredient Talazoparib (breast cancer, BRCA1/2 mutation, HER2-). Benefit assessment according to Section 35a SGB V. Published online September 1, 2020.
4. Gemeinsamer Bundesausschuss (G-BA). Benefit assessment procedure for the active ingredient osimertinib (new indication: non-small cell lung cancer, first-line). Benefit assessment according to Section 35a SGB V. Published online October 15, 2018. https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/377. Accessed August 4, 2025.
5. Gemeinsamer Bundesausschuss (G-BA). Sacituzumab Govitecan (Mammakarzinom, HR+, HER2-, mind. 3 Vortherapien) – Nutzenbewertung gemäß §35a SGB V. [Sacituzumab govitecan (breast cancer, HR+, HER2-, at least 3 prior therapies) – Benefit assessment according to §35a Social Code Book V.] Published online May 19, 2022 at https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/986. Accessed January 14, 2025.

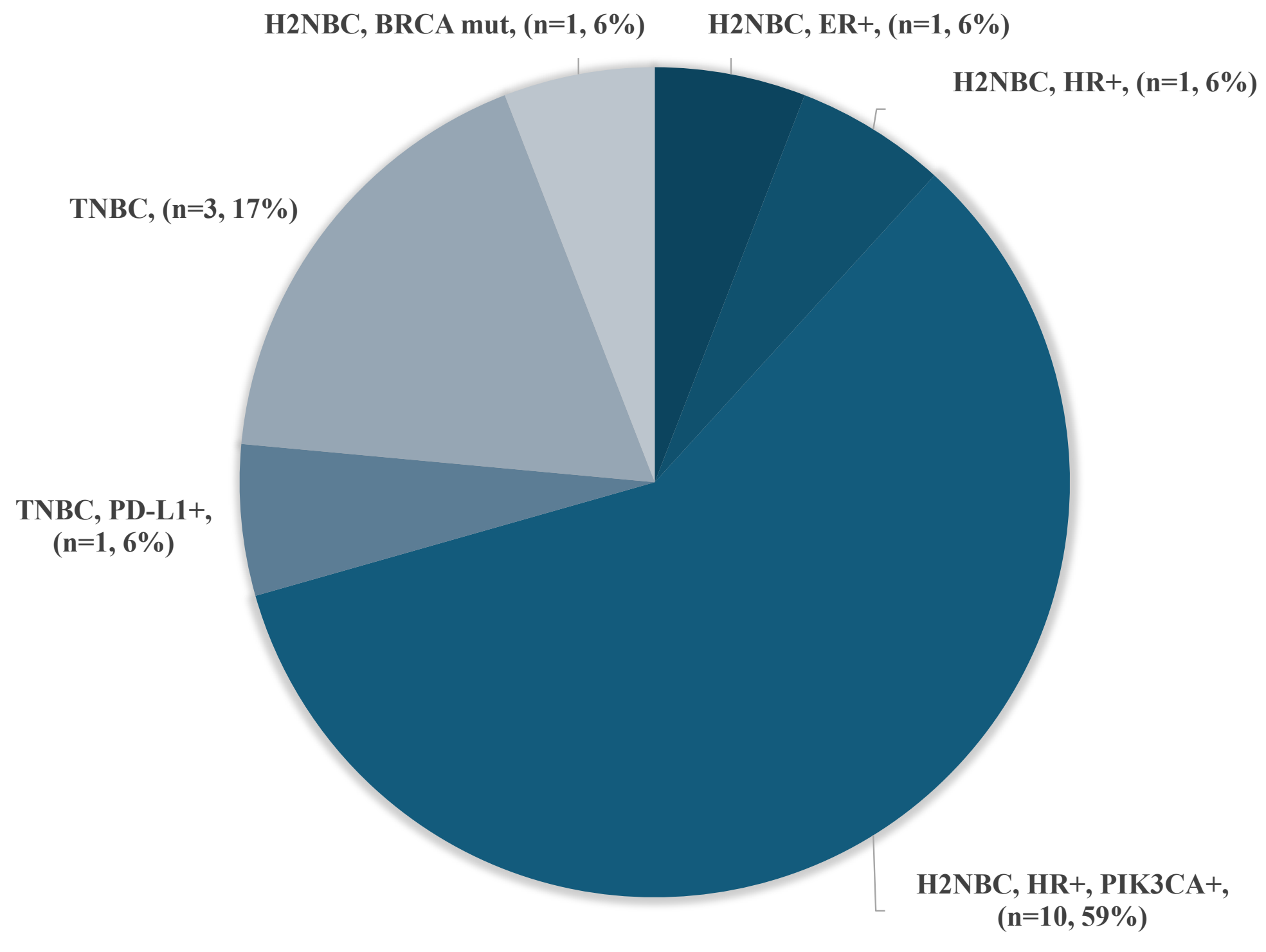

Figure 1. H2NBC assessments (n=17) by cancer subtype

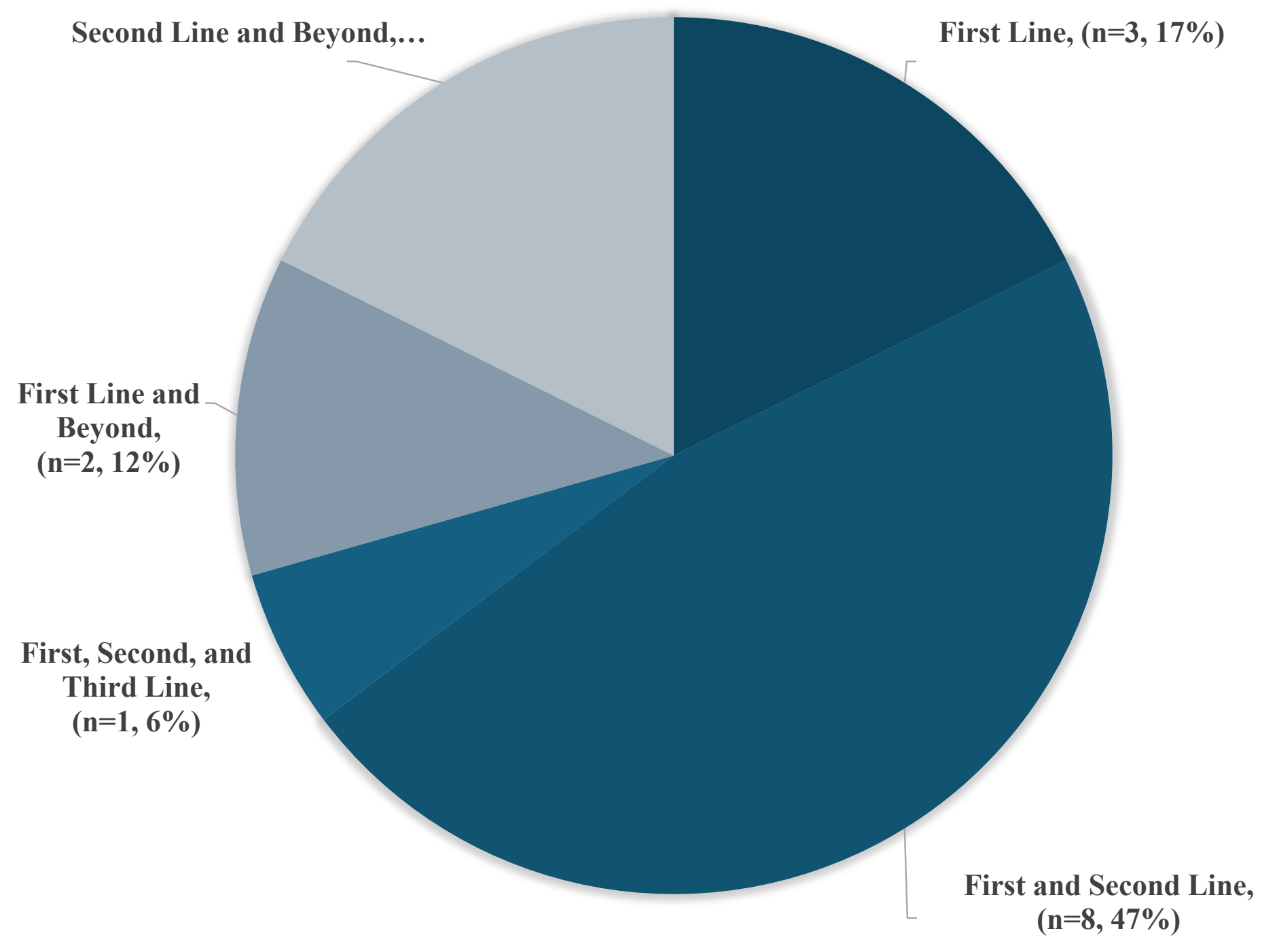

Figure 2. H2NBC assessments (n=17) by the line of therapy

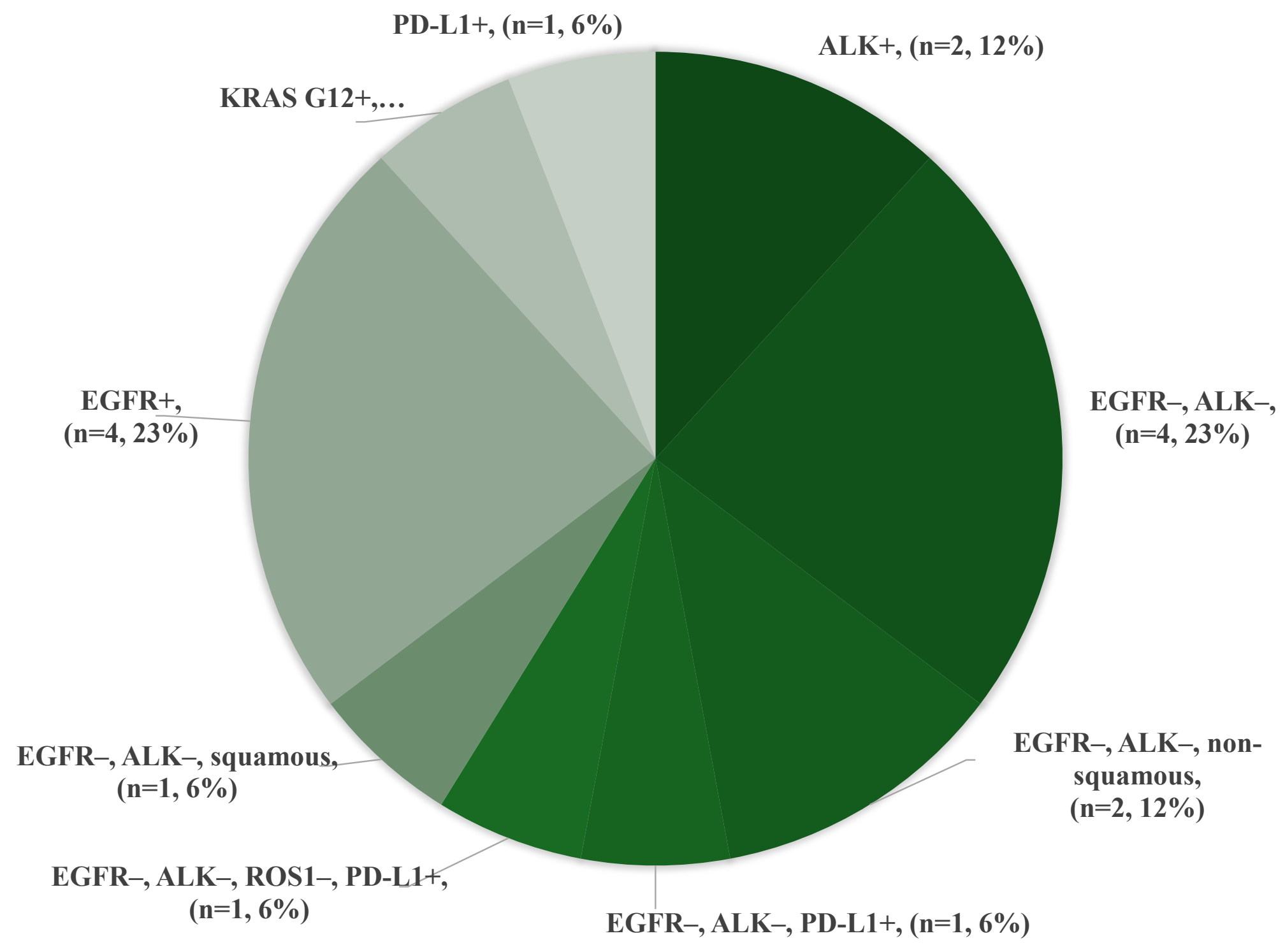

Figure 4. NSCLC assessments (n=17) by cancer subtype

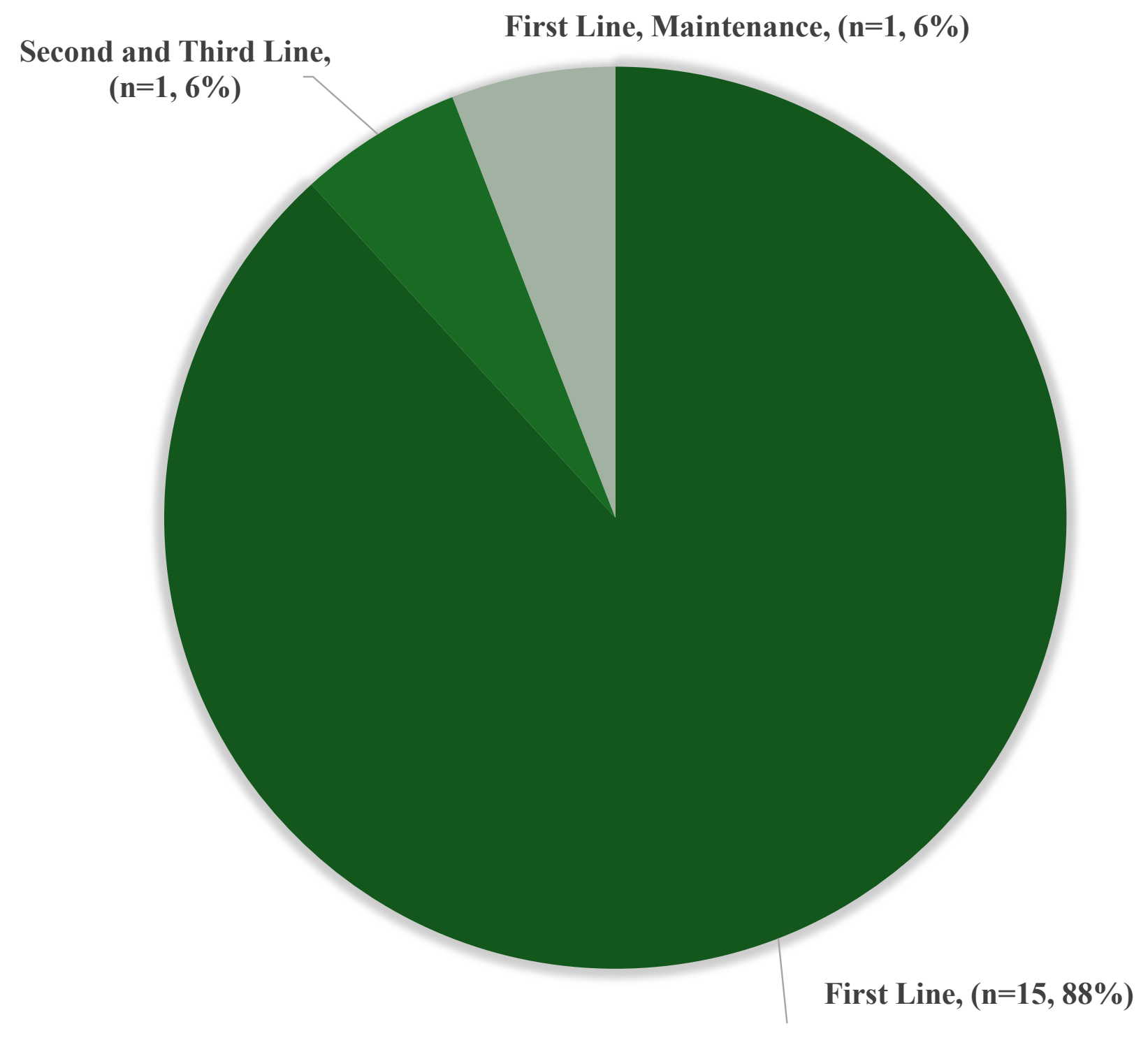

Figure 3. NSCLC assessments (n=17) by the line of therapy

Types of Critique	Illustrative Examples of Critique	Effect on the Assessment
Availability of data:		
Absence of relevant PRO data (H2NBC n=5; NSCLC n=9)	▪ Absence of data necessary to ensure a fair comparison between the treatment arms: ▪ questionnaire response rates ▪ information on treatment and observation periods ▪ Kaplan-Meier curves to accompany the time-to-event analyses ▪ information on treatment and observation periods ▪ Absence of PRO data in at least one clinical trial in the submission ▪ Absence of MMRM analyses for EQ-5D VAS ▪ Absence of collected QLQ-LC13 data in submission ▪ Lack of usable data on HRQoL ▪ Absence of the subgroup analyses for the PRO-based endpoints	▪ In all instances, PRO data were not included in the analyses
AIC analyses not possible due to difference in study design (H2NBC n=0; NSCLC n=5)	▪ Two studies used in AIC had different design ▪ Two comparator studies, one had high risk of bias due to open-label trial design, and another did not include PRO data	▪ In all instances, PRO data were not included in the analyses
Clinical trial design:		
Open-label trial design (H2NBC n=7; NSCLC n=9)	▪ PRO data were assigned high risk of bias due to open-label trial design ▪ In a double-blinded study because investigators were allowed to unblind patients to the treatment after the disease progression, creating a possibility for unblinding ▪ Blinding was assumed impossible to maintain because of the known AEs of the investigational	▪ In all instances, the certainty of evidence based was reduced.
Criticism of observation period (H2NBC n=12; NSCLC n=6)	▪ Incomplete observations for potentially informative reasons, resulting in different length of the follow-up period between arms ▪ The company changed dosing of investigational drug during clinical trial ▪ PRO collection was 30 days post-progression, which resulted in misbalanced observation periods between study arms ▪ The schedule of assessment did not match the schedule of chemotherapy administration for some timepoints. ▪ The discontinuation rate was misbalanced between arms for potentially informative reasons ▪ PRO assessment (until 1st progression) was shorter than drug administration (until 2nd progression) ▪ Insufficiently long observation period: 30- and 90-days post progression ▪ Observation time for PROs was systematically shortened (30 days after last drug administration) to make reliable statement about entire study period until death, necessary to record endpoints over the entire period ▪ Selective extended follow-up in the active-treatment arm: only subjects who receive active treatment as follow-up therapy after progression were followed	The data were rejected in the following cases: ▪ Schedule of assessment did not match the schedule of chemotherapy administration for some timepoints ▪ PRO assessment (until 1st progression) was shorter than drug administration (until 2nd progression) ▪ In one instance of PRO data collected insufficiently long (30 days post-progression) ▪ The short observation period was insufficient to support the definition of "definitive deterioration" In the remaining instances, the certainty of the evidence base was reduced
High level of missing data (H2NBC n=3; NSCLC n=7)	▪ Substantial decrease in response rates to questionnaires that differed between treatment arms ▪ A high proportion of patients (> 10 percentage points) were not included in the analysis due to missing values at baseline ▪ High level of missing data overall	IQWiG reduced the certainty of the evidence base in all cases except one when missing data were explained by death
Discrepancies between study arms (H2NBC n=2; NSCLC n=2)	▪ Treatment arms differed in the proportion of patients included in the analyses ▪ Large difference (about 12 percentage points) between the study arms in the proportion of patients who discontinued the study before the first administration of the study medication. This results in a high difference (≥5 percentage points) in the proportion of patients not included in the evaluation between the treatment arm	▪ When discrepancy between study arms exceeded 15 percentage points, the endpoints were rejected ▪ When the discrepancy did not exceed 15 percentage points, the endpoints were accepted, but the certainty of the evidence base was reduced
Shortcomings of Analyses:		
Discrepancy with the pre-defined plan (H2NBC n=1; NSCLC n=3)	▪ The cutoff for PRO endpoints was non-planned and was earlier than the cutoff for the mortality data ▪ The submitted analysis was not pre-defined ▪ The executed analyses did not match the ones described in the SAP, without an explanation of why ▪ The planned analyses were not submitted	▪ In all instances, PRO data were rejected
Data censoring issues (H2NBC n=4; NSCLC n=4)	▪ High censoring rates for potentially informative reasons ▪ Only participants with baseline scores ≤90 points for the function and GHS/QoL scales and ≥10 points for symptom scales were included in EORTC QLQ-C30 and QLQ-LC13 TTD analyses ▪ QLQ-BR23 score for "sexual enjoyment" was estimated only for patients who were sexually active at baseline ▪ MMRM analysis included only patients who had at least 2 measured values, which differed by 15 percentage points or more between treatment arms	▪ Data were rejected when censoring was based on baseline values or created discrepancy between arms that exceeded 15 percentage points ▪ In other instances, the certainty of the evidence base was reduced
Insufficient definition of endpoints (H2NBC n=4; NSCLC n=2)	▪ From the description of time to definitive deterioration endpoint, it was unclear whether a subsequent improvement referred exclusively to the next subsequent recording or to all further subsequent recordings, how exactly the subsequent improvement was operationalized, and how many patients were included in the analysis ▪ Time to deterioration endpoints were rejected because the threshold for detereoration was not specified ▪ In TTDD analysis, it was unclear whether patients who did not have follow-up surveys after observation of first deterioration were classified as patients with events or censored ▪ The information on the actual observation periods of the outcomes on morbidity, HRQoL and side effects in the company's dossier is not comprehensible	▪ All endpoints with vague definitions were rejected
Disagreement with endpoint definition (H2NBC n=12; NSCLC n=7)	▪ In the time to detereoration analyses, the detereoration was based on an unacceptable threshold, e.g.: ▪ 7- or 10-points for EQ-5D VAS ▪ ≥2 points for BPI-SF composed scores of pain intensity and pain interference ▪ Time to detereoration analyses defined detereoration as score worsening or death while death was reflected in the overall survival ▪ Time to definitive deterioration was defined as: deterioration by at least the response threshold without subsequent improvement, or deterioration by at least the response threshold and no subsequent values ▪ Time to improvement was rejected because time to worsening was considered more adequate in advanced-stage disease ▪ An endpoint of symptomatic progression (based on EORTC QLQ-C30 symptoms worsening) was rejected because IQWiG disagreed with the endpoint definition (multiple reasons, including the selection of symptoms, possibility of symptoms worsening because of something else, besides progression, and because 7-week period not adequately justified ▪ The number of expected responses was the number of patients who could theoretically complete the questionnaire (i.e. living patients, patients without lost to follow-up, etc.) instead of the number of all patients who have not died by this point in time ▪ For time to detereoration in pain, the company did not present a separate analysis for each component of the endpoint, which was defined as either increase in mBPI-SF worst pain by ≥ 2 points and the increase in analgesics use	▪ IQWiG rejected the endpoints, the definitions of which it disagreed with
Relevance Concerns		
The absence of justification for selecting specific PRO concepts (H2NBC n=2; NSCLC n=1)	▪ Lack of rationale for the selection of the specific PRO-CTCAE concepts for evaluation	▪ PRO-CTCAE analyses were rejected
Symptoms rated as non-serious, non-severe in the absence of evidence of degree of seriousness and severity (H2NBC n=9; NSCLC n=2)	▪ For statistically significant results based on the EORTC QLQ-C30 symptom scales, the sponsor did not provide information on the classification of severity or seriousness for these outcomes	▪ IQWiG assigned "non-serious" status to the symptoms that did not have the level of seriousness and severity specified

Presented at ISPOR Europe 2025 | 10 November 2025 | Glasgow, Scotland, UK