Stellungnahme zum Entwurf der Version 6.0 des IQWiG-Papiers
„Allgemeine Methoden“
durch die gemeinsame Präsidiumskommission
„Methodenaspekte in der Arbeit des IQWiG und IQTIG“ der GMDS und IBS-DR

Im „Allgemeinen Methodenpapier“ werden die gesetzlichen und wissenschaftlichen Grundlagen der Arbeit am Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen (IQWiG) erläutert. Die derzeit gültige Fassung ist die Version 5.0 vom 10.07.2017. Am 5. Dezember 2019 hat nun das IQWiG einen Entwurf der neuen Version 6.0 vorgelegt. In der Version 6.0 wurden im Vergleich zur Version 5.0 neben redaktionellen Modifikationen auch inhaltliche Änderungen durchgeführt. Grundlegende Überarbeitungen erfolgten in den Kapiteln zur *Nutzenbewertung medizinischer Interventionen* (Kapitel 3), *Bewertung der Versorgung* (Kapitel 5), *HTA-Berichte* (Kapitel 6) und *Informationsbewertung* (Kapitel 9). Aus biometrischer Sicht sind insbesondere einige Änderungen in den Kapiteln 3 und 9 relevant. Auf diese gehen wir unten näher ein.

Die Kommission begrüßt die regelmäßige Überarbeitung der „Allgemeinen Methoden“. Die stetige Fortentwicklung relevanter Methoden macht dies unerlässlich. Wie aus den spezifischen Kommentaren unten deutlich wird, gibt es einige Abschnitte, die einer Aktualisierung bedürfen. Die Kommission ist offen für eine fachliche Diskussion mit dem IQWiG in Bezug auf geeignete Methoden und unterstützt das IQWiG in dieser Hinsicht gerne.

Die Kommission hat zu den folgenden speziellen Punkten des Papiers Kommentierungen durchgeführt:

Abschnitt 3.3.3 Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V
Der Abschnitt 3.3.3 widmet sich der Nutzenbewertung von Arzneimitteln gemäß § 35a SGB V. Unter Punkt D) werden Schwellenwerte für die Ausmaßbestimmung für stetige oder quasistetige Zielgrößen mit jeweils vorliegenden standardisierten Mittelwertdifferenzen (SMD) behandelt. Diese sind in der Tabelle 6 zusammengefasst. Aus Sicht der Kommission ist dies eine sinnvolle Erweiterung. Allerdings ist der folgende Satz unverständlich: „Man kann zeigen, dass die für ein bestimmtes Ausmaß wünschenswerte Effektstärke ungefähr halbiert werden muss, um den entsprechenden Schwellenwert abzuleiten.“ Hier wünscht sich die Kommission eine nachvollziehbare Begründung. Dies könnte z.B. durch eine Erweiterung des Anhangs A im Hinblick auf SMD geschehen.

Abschnitt 9.3.8 Metaanalysen
Im Abschnitt 9.3.8 wird das Vorgehen bei Metaanalysen detailliert. Unter anderem wird hier festgestellt, dass bei Vorliegen einer ausreichenden Anzahl von Studien (interpretiert als wenigstens fünf Studien) die Knapp-Hartung-Methode für Metaanalysen mit zufälligen Effekten zur Anwendung kommen sollte. Bei der Knapp-Hartung-Methode werden zur Konstruktion von Konfidenzintervallen t-Quantile und ein skalierter Varianzschätzer verwendet. Die Skalierung des Varianzschätzers kann in seltenen Fällen dazu führen, dass das Konfidenzintervall beruhend auf der Knapp-Hartung-Methode kürzer ist als die Intervalle, die z.B. aus einer Metaanalyse mit festen Effekten oder eine Metaanalyse

der DerSimonian-Laird-Methode resultieren. Hier kann man sich mit einer einfachen ad hoc Korrektur behelfen, indem der Skalierungsfaktor für den Varianzschätzer auf einen Bereich von 1 oder größer beschränkt wird. Diese ad hoc Korrektur wurde von Hartung und Knapp im Kontext mit Metaregressionen beschrieben (Referenz [421]). Die Kommission hat keine Einwände in Bezug auf die Verwendung dieser ad hoc Korrektur. Allerdings ist die Kommission der Meinung das einige Formulierungen in dem Absatz beginnend mit „Bei Anwendung der Knapp-Hartung-Methode für Metaanalysen mit zufälligen Effekten“ (Seite 183) potentiell irreführend sind. So ist es z.B. sachlich nicht richtig, dass Knapp und Hartung in dem zitierten Aufsatz die Anwendung dieser Korrektur für die beschriebene Situation vorgeschlagen haben. Zudem ist der Hinweis auf die „Knapp-Hartung-Methode mit Varianzkorrektur“ nicht eindeutig, da diese Methode verglichen z.B. mit der DerSimonian-Laird-Methode immer eine Varianzkorrektur beinhaltet. Dies sollte eindeutiger formuliert werden, z.B. durch die Verwendung eines Begriffs wie „ad hoc Korrektur“ mit einem eindeutigen Bezug auf die Literatur.

Unter Punkt D) im Abschnitt 9.3.8 finden sich Erläuterungen zu Metaanalysen von Studien zur diagnostischen Güte. Aus Sicht der Kommission sollte dieser Punkt dringend aktualisiert werden, da einige methodische Entwicklungen der letzten Jahre sich hier leider nicht wiederfinden und sich somit Widersprüche zu Empfehlungen z.B. der Cochrane Collaboration ergeben. Dies betrifft insbesondere die Berechnung von summarischen ROC (SROC)-Kurven, sowie die Verwendung von getrennten Modellen zur Schätzung der Sensitivität und Spezifität.

Für die Berechnung von SROC-Kurven ist eine entsprechende Datengrundlage notwendig, die Angaben zu allen in den Einzelstudien geschätzten Sensitivitäten und Spezifitäten enthält. Die unter Punkt D) genannten statistischen Verfahren erlauben zwar Metaregressionen, gehen aber davon aus, dass jeweils ein diagnostischer Schwellenwert pro Studie selektiert wird. Damit sind sie für die komplexere Situation von mehreren Schwellenwerten auf Studienebene nicht geeignet. Hinzu kommt, dass die resultierenden SROC-Kurven nur schwer interpretierbar sind (Arends et al, 2008). Hierfür gibt es neuere statistische Verfahren, die zitiert werden sollten (Steinhauser et al, 2016; Hoyer et al, 2018; Jones et al, 2019). Weiterhin wird unter Punkt D) aufgeführt, dass zur metaanalytischen Zusammenfassung getrennte Modelle für Sensitivität und Spezifität verwendet werden können. Der Hinweis auf diese Möglichkeit ist insofern irreführend, als dass er im Gegensatz zu den Empfehlungen der Cochrane Collaboration steht, welche die Verwendung von bivariaten bzw. hierarchischen Modellen nahelegt (Macaskill et al, 2010).

Abschnitt 9.3.12 Umgang mit unvollständigen Daten

Im Abschnitt 9.3.12 wir der Umgang mit unvollständigen Daten behandelt. Hierbei werden konkrete Verfahrensregeln festgelegt. Diese sehen vor, dass Studien, in denen weniger als 70% der eingeschlossenen Patienten (d.h. der Intent-to-Treat-Population) zur Auswertung beitragen oder der Anteil der in die Auswertung eingeschlossenen Patienten zwischen den Behandlungsgruppen sich um mindestens 15 Prozentpunkte unterscheiden, nicht berücksichtigt werden. Aus Sicht der Kommission fehlt eine Rationale für die festgelegten Grenzen von 70% und 15 Prozentpunkten. Zudem werden auch keine belastbaren Referenzen genannt. Die Kommission empfiehlt darüber hinaus, Formulierungen zu verwenden, die Ausnahmen zulassen. Zum Beispiel könnte sich eine Situation ergeben, in der unter Standardtherapie aufgrund von unerwünschten Ereignissen sehr viele Patienten vorzeitig die Studie beenden, während dies im experimentellen Arm nicht der Fall ist. Dann würde hier u.U. die experimentelle Therapie zu Unrecht bestraft. Diese Situationen sind in der Darstellung im Nutzendossier bezüglich ihrer Auswirkungen auf die Effektschätzer und den zugrundeliegenden Estimand zu diskutieren. Insbesondere ist aus Sicht der Kommission eine grundlegende Überarbeitung

dieses Abschnitts dringend geboten, da dieser der aktuellen Estimand-Diskussion und neueren Literatur nicht gerecht wird (siehe z.B. Referenz [699]).

Abschnitt 9.3.13 Umgang mit unterschiedlichen mittleren Beobachtungsdauern

Der Abschnitt 9.3.13 behandelt den Umgang mit unterschiedlichen mittleren Beobachtungsdauern. Die Kommission begrüßt, dass dieser Abschnitt um neuere Entwicklungen bei der Analyse von unerwünschten Ereignissen ergänzt wurde. So ist nun z.B. ein Verweis auf den kürzlich erschienenen Artikel von Unkel et al. enthalten (Referenz [699]), in dem Ergebnisse einer gemeinsamen Projektgruppe von GMDS und IBS-DR zur Auswertung von unerwünschten Ereignissen bei variierenden Nachbeobachtungszeiten und konkurrierenden Ereignissen zusammengetragen wurden. Allerdings sieht die Kommission als problematisch an, dass hier lediglich die Situation unterschiedlicher mittlerer Beobachtungszeiten zwischen Behandlungsgruppen behandelt wird. Auch im Falle unterschiedlicher individueller Beobachtungszeiten innerhalb einer Behandlungsgruppe können ungeeignete Analysemethoden zu verzerrten Schätzungen der Wahrscheinlichkeit für ein unerwünschtes Ereignis innerhalb der Behandlungsgruppe führen. Dies kann einerseits zu einer falschen Beurteilung von Sicherheitsrisiken innerhalb der Behandlungsgruppen und andererseits ebenfalls zu Verzerrungen bei der Betrachtung von Gruppenunterschieden führen. Des Weiteren sollte aus Sicht der Kommission unbedingt betont werden, dass konkurrierende Ereignisse (wie Tod oder vorzeitiges Ende der Nachbeobachtung bzgl. unerwünschter Ereignisse) in der Analyse durch adäquate Methoden berücksichtigt werden müssen. Ein globaler Verweis auf Techniken der Überlebenszeitanalyse ist nicht ausreichend. Dieser globale Hinweis ist sogar irreführend, da der häufig verwendete Kaplan-Meier-Schätzer (d.h. 1 – Kaplan-Meier-Kurve zum entsprechenden Zeitpunkt) die Wahrscheinlichkeit eines unerwünschten Ereignisses bei Vorliegen konkurrierender Ereignisse überschätzt. Darüber hinaus sieht die Kommission als besonders problematisch an, dass in der Dokumentvorlage für das Dossier zur Nutzenbewertung (Version vom 21.02.2019) explizit die Erstellung von Kaplan-Meier-Kurven gefordert wird. Es sollte darauf hingewiesen werden, dass bei Vorliegen von konkurrierenden Ereignissen der Aalen-Johansen-Schätzer für die Wahrscheinlichkeit eines unerwünschten Ereignisses zu verwenden ist.

Referenzen

- Allignol A, Beyersmann J, Schmoor C (2016) Statistical issues in the analysis of adverse events in time-to-event data. Pharmaceutical Statistics 15: 297-305
- Arends LR, Hamza TH, van Houwelingen JC, Heijenbrok-Kal MH, Hunink MGM, Stijnen T (2008) Bivariate Random Effects Meta-Analysis of ROC Curves. Medical Decision Making 28: 621-638
- Steinhauser S, Schumacher M, Rücker G (2016) Modelling multiple thresholds in meta-analysis of diagnostic test accuracy studies. BMC Medical Research Methodology 16: 97
- Hoyer A, Hirt S, Kuss O (2018) Meta-analysis of full ROC curves using bivariate time-to-event models for interval-censored data. Research Synthesis Methods 9: 62-72
- Jones HE, Gatsonsis CA, Trikalinos TA, Welton NJ, Ades AE (2019) Quantifying how diagnostic test accuracy depends on threshold in a meta-analysis. Statistics in Medicine 38: 4789-4803
- Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/