

Anton Gies

Dr. sc. hum.

Evaluation of quantitative fecal immunochemical tests for detection of occult blood in stool for colorectal cancer screening

Fach/Einrichtung: DKFZ (Deutsches Krebsforschungszentrum)

Doktorvater: Prof. Dr. med. Hermann Brenner

Eine Vielzahl von unterschiedlichen quantitativen immunologischer Stuhltests zum Nachweis von humanem Hämoglobin wird zum Darmkrebsscreening eingesetzt. Es ist jedoch unklar, ob echte Unterschiede in Bezug auf relevante Schlüsselfaktoren, wie z. B. die diagnostische Güte, zwischen den Tests bestehen. Das Ziel dieser Dissertation war es, mögliche Gründe für die berichteten Unterschiede in der Sensitivität und Spezifität verschiedener Tests zu untersuchen, sowie die diagnostische Güte (inklusive der Bewertung, ob die Testleistung durch Kombination verschiedener Tests miteinander oder in Kombination mit Transferrin verbessert werden kann), die Probenstabilität und verschiedene Techniken zur Stuhlprobeentnahme von einer großen Anzahl an unterschiedlichen Tests zu bewerten und zu vergleichen.

Die Ergebnisse des systematischen Reviews ergaben große Unterschiede zwischen den eingeschlossenen Studien (n=22) in den Studienpopulation (z.B. unterschiedliche Alters-, Geschlechts- und Krebsstadien-Verteilungen), prä-analytischen Testbedingungen (Lagerzeit und Temperatur) und unterschiedlich angewandten Cutoff-Werten (Bereich: 2-82 $\mu\text{g/g}$), was zu unterschiedlichen Spezifitäten (Bereich: 82-99%) und Sensitivitäten zum Nachweis von Darmkrebs (Bereich: 25-100%) und fortgeschrittenen Neoplasien (Darmkrebs oder fortgeschrittenes Adenom; Bereich: 9-60%) führte. Dabei korrelierte die Positivitätsrate fast perfekt mit der Spezifität (Koeffizient: 0,927) und wies eine sehr starke Vorhersagekraft für die Sensitivität auf (Koeffizient: 0,836), während der Cutoff-Wert alleine nur eine moderate Korrelation mit der Spezifität (Koeffizient: 0,605) und die Sensitivität (Koeffizient: 0,526) aufwies. Dies deutet darauf hin, dass verschiedene Tests bei entsprechender Anpassung der Cutoff-Werte eine vergleichbare diagnostische Güte haben könnten. Aufgrund der großen Unterschiede zwischen den Studiendesigns konnte dies jedoch nicht sicher bewertet werden. Daher wurde eine große Anzahl verschiedener Tests von verschiedenen Herstellern parallel unter den gleichen prä-analytischen und analytischen Testbedingungen evaluiert, wobei genau

dieselben Stuhlproben von derselben Kohorte an Teilnehmern verwendet wurden, die sich für Vorsorge-Koloskopie vorstellten (Goldstandard=Referenzstandard). Bei Auswertung der Tests bei ihren ursprünglichen Cutoff-Werten (Bereich: 2-17 µg/g) wurden teilweise große Unterschiede in der Sensitivität (Bereich: 22-46%; für fortgeschrittener Neoplasien) und Spezifität (Bereich: 86-98%) beobachtet. Selbst die Vereinheitlichung der Cutoff-Werte lieferte keine vergleichbaren Ergebnisse. Passte man die Cutoff-Werte jedoch so an, dass alle Tests dieselbe Spezifität lieferten (oder Positivitätsrate; die nahezu perfekt miteinander korrelieren), wurden sehr ähnliche Sensitivitäten bei nahezu identische Positivitätsraten (bzw. Spezifitäten) beobachtet. Bei einer Spezifität von beispielsweise 97% lag die Sensitivität für fortgeschrittene Neoplasien zwischen 21% und 24%, bei nahezu identischen Positivitätsraten von 5,8% bis 6,1%. Darüber hinaus fanden sich Hinweise auf einige testspezifische Unterschiede in der diagnostischen Güte nach Geschlecht, Alter, Lokalisation und Krebs-Stadium. Kombinationen verschiedener Tests miteinander oder die Kombination mit Transferrin führten zu keiner Verbesserung der diagnostischen Güte im Vergleich zur Anwendung eines einzelnen Stuhltests auf Hämoglobin.

Verschiedene Testhersteller verwenden unterschiedliche Stabilisationslösungen, um fäkales Hämoglobin bis zur Testanalyse zu stabilisieren. Um zu beurteilen, ob dies zu unterschiedlichen Hämoglobinstabilitäten führt, wurden 10 unterschiedlichen Tests simultan, innerhalb derselben Studienpopulation und unter kontrollierten prä-analytischen Lagerbedingungen, bewertet und verglichen. Bei 5°C und 20°C Lagertemperatur zeigten fast alle Tests ziemlich konstante Messergebnisse und Positivitätsraten innerhalb des Beobachtungszeitraums von sieben Tagen, während bei 35°C große Unterschiede zwischen den verschiedenen Tests gefunden wurden. Insgesamt präsentierten nur CAREprime Hb und OC-Sensor ausreichend stabile Ergebnisse ohne einen statistisch signifikanten Abfall der gemessenen Hämoglobinkonzentration.

Obwohl die Instruktionen zur Stuhlprobeentnahme unter den Testherstellern sehr unterschiedlich waren, konnten diese durch drei Schlüsselemente beschrieben werden: (i) Mehrfache Insertionen des Probeentnahmestäbchens in die Stuhlprobe, (ii) optische Inspektion der Kerben des Stäbchens auf vollständigen Befüllung mit Stuhl, und (iii) einmaliger Transfer der gesammelten Stuhlprobe in das Teströhrchen. Verstöße gegen die empfohlenen Instruktionen basierend auf den Schlüsselementen (i) und (ii) ergaben niedrigere Messwerte in der Hämoglobinkonzentration, wohingegen der Verstoß gegen das dritte Schlüsselement wesentlich höhere Messwerte ergab. Nichtsdestotrotz blieben die Sensitivitäten und Spezifitäten bei Verstößen gegen die ersten beiden Elemente weitgehend

konstant. Beim Verstoß gegen das dritte Schlüsselement (Transfer von mehr Stuhlmenge in das Röhrchen als empfohlen) wurden sogar wesentlich höhere Sensitivitäten bei hohen Spezifitätsniveaus erreicht (in den meisten Fällen über 90%). Diese Ergebnisse sprechen für eine hohe Robustheit der immunologischen Stuhltests gegenüber Fehlern in der Probeentnahme. Dennoch sollten die Instruktionen zur Probeentnahme klar, einfach und leicht umsetzbar sein, um Fehler in der Entnahme und daraus resultierende mögliche Variationen der Messergebnisse zu minimieren.

Zusammenfassend lässt sich sagen, dass eine große Anzahl unterschiedlicher quantitativer immunologischer Stuhltests ein ähnlich hohes Potenzial für die Erkennung von Darmkrebs und fortgeschrittenen Adenomen bei hoher Spezifität aufweisen. Sensitivitäten und Spezifitäten variierten jedoch stark, wenn die von den Herstellern empfohlenen Grenzwerte für ein positives Testergebnis angewendet wurden. Daher sollten Screening-Programme in Betracht ziehen, den voreingestellten Grenzwert des Testherstellers auf ein beabsichtigtes Spezifitäts-, oder Positivitätsniveau (das nahezu perfekt mit der Spezifität korreliert und den Bedarf an Follow-up-Koloskopien determiniert) anzupassen, anstatt einfach den voreingestellten Cutoff-Wert anzuwenden und daran ggf. festzuhalten. Es sollten weitere testspezifische Merkmale in Bezug auf die diagnostische Güte nach Geschlecht, Alter, Lokalisation und CRC-Stadium sowie testspezifische Charakteristika der Probenstabilität (zusätzlich zu den Merkmalen wie Kosten pro Test, Benutzerfreundlichkeit und Weite des analytischen Messbereichs) bei der Auswahl eines Tests für das Screening berücksichtigt werden. Obwohl die Tests relativ robust gegenüber Verstößen in der Probeentnahme zu sein scheinen, sollten die Instruktionen klar, einfach und leicht zu befolgen sein, um konstante Messwerte und diagnostische Güte zu erzielen. Das Potential einer Verbesserung der diagnostischen Wertigkeit durch Füllen von mehr Stuhl in das Teströhrchen als derzeit praktiziert sollte in weiteren Studien untersucht werden.