EINE DATENVERARBEITUNGSPIPELINE ALS BASIS FÜR KREBSREGISTERÜBERGREIFENDE DATENANALYSEN UND KI-ANWENDUNGEN (AI-CARE)

Dr. Markus Sauerberg, Hamburgisches Krebsregister

Krebsregisterübergreifende Analysen sind mit den praktischen Herausforderungen einer gewissen Datenheterogenität verbunden. Zwar melden Kliniken und Praxen ihre Informationen im Format des bundeseinheitlichen onkologischen Basisdatensatzes an die flächendeckenden Krebsregister Deutschlands, wo die weitere Verarbeitung für Auswertungen jedoch unterschiedlich erfolgt. Ein registerübergreifender harmonisierter Datensatz fördert die Aussagekraft statistischer Analysen erheblich, z.B. betr. seltene Krebsentitäten oder besonderer onkologischer Therapien. Auch für moderne Methoden, wie dem maschinellen Lernen, werden große qualitätsgesicherte Datensätze benötigt.

Für diese Zwecke haben wir innerhalb des AI-CARE Projekts eine nutzerfreundliche Datenverarbeitungspipeline mit der Software R erstellt, die Krebsregisterdaten aus verschiedenen Bundesländern in einen harmonisierten Gesamtdatensatz überführt. Die Pipeline basiert auf dem Lieferdatensatzformat des Zentrums für Krebsregisterdaten (ZfKD), und ist im Rahmen von AI-CARE auf die vier Entitäten Brustkrebs, Lungenkrebs, Schilddrüsenkrebs und das Non-Hodgkin-Lymphom beschränkt.

Jeder gemeldete Wert wird durch die Datenverarbeitungspipeline mithilfe von Referenztabellen auf seine Gültigkeit geprüft. Ungültige Werte werden entweder in gültige Werte übersetzt oder entfernt. Momentan beinhaltet der durch die Datenverarbeitungspipeline bereinigte Gesamtdatensatz ca. 726 000 Tumore von 705 000 Patientinnen und Patienten aus elf Bundesländern.

Die Datenverarbeitungspipeline bietet die Möglichkeit, einen bundesweit einheitlichen und plausibilisierten Auswertungsdatensatz erstellen zu können. Der Datensatz wird beispielsweise aktuell verwendet, um mit KI-Verfahren Faktoren zu ermitteln, die einen Einfluss auf die Überlebensprognose bei Lungenkrebs aufweisen.