„ES FEHLT AN GLOBALEN DATENSÄTZEN, UM ZU BESTIMMTEN KRANKHEITEN EINE KI ZU TRAINIEREN … NIEMAND HAT SICH BISHER DER AUFGABE ANGENOMMEN, GLOBALE DATENSÄTZE ZUSAMMENZUSTELLEN."

Bart de Witte

Im vergangenen Jahr geriet der Techriese Google in Wissenschaftskreisen schwer in die Kritik. Das Google-Health-Team des US-Unternehmens hatte in der Fachzeitschrift „Nature" einen viel beachteten Beitrag veröffentlicht. Googles künstliche Intelligenz (KI) DeepMind habe bei einer Studie bei der Erkennung von Brustkrebs besser abgeschnitten als menschliche Experten, hieß es in der Veröffentlichung. In einem offenen Brief kritisierten Wissenschaftler namhafter Hochschulen jedoch, dass zu wenige Informationen zum Algorithmus und den genauen Methoden preisgegeben worden seien – und somit das Vorgehen nicht überprüfbar sei, wie es gute Wissenschaft verlange. Google wollte mit Verweis auf den Einsatz bestimmter firmeneigener Tools keine weiteren Informationen zur Verfügung stellen.

Für Bart de Witte war diese Auseinandersetzung in der Wissenschaftscommunity ein Aha-Moment. „Ich habe gleich gedacht: Das ist die europäische Chance!", sagt de Witte, der früher bei IBM im deutschsprachigen Raum für das Thema Digital Health verantwortlich war. Googles Verhalten weist seiner Ansicht nach den Weg, wie europäische Unternehmen aus dem Gesundheitswesen im Wettbewerb mit den großen Techunternehmen bei KI-Themen bestehen können. Zwar hätten Google, Amazon und Co. sehr viel Kapital, um Daten zu sammeln und damit ihre KI-Lösungen für den Gesundheitssektor weiterzuentwickeln. Die wesentlich kleineren europäischen Unternehmen mit entschieden weniger Kapital hätten in diesem Wettrennen um die besten KI-Lösungen wegen der zudem kleineren Märkte keine Chance. De Witte sieht aber einen Ausweg: Open-Source-Standards. Wenn in Europa jeder seine Algorithmen teilen müsste, wäre etwa Google außen vor. „Das wäre eine große Chance für die Gesellschaft und auch die Industrie in Europa mit ihren 35.000 kleinen und mittelständischen Unternehmen im Medtech-Sektor", sagt de Witte.

Sein Vorschlag ist sehr weitgehend. Viele Unternehmerinnen und Unternehmer dürften sich nicht recht mit seiner Idee anfreunden. Denn wer gibt schon gerne kostenlos sein geistiges Eigentum her? Unabhängig davon, ob man de Wittes Ansatz gut findet, wirft der Vorschlag des Digital-Health-Experten jedoch ein Schlaglicht auf eine zentrale Herausforderung, vor der Unternehmen aus dem Gesundheitswesen stehen: Woher bekommen Unternehmen ausreichend hochwertige, strukturierte Daten, um eine KI zu trainieren? Denn daran mangelt es.

Text

„Es fehlt an globalen Datensätzen, um zu bestimmten Krankheiten eine KI zu trainieren", sagt de Witte. Es gibt zwar frei zugängliche Datensätze, wie etwa zu Hautkrebs. Doch diese seien verzerrt und nicht repräsentativ, weil nicht alle Hauttypen abgedeckt seien, so de Witte. „Niemand hat sich bisher der Aufgabe angenommen, globale Datensätze zusammenzustellen", sagt der Digital-Health-Experte. Das will er nun mit der Hippo AI Stiftung ändern.

De Wittes großes Ziel ist es, jedem Menschen auf dem Globus den Zugang zur gleichen Diagnostik zu ermöglichen. Mit seiner Stiftung arbeitet er derzeit zum einen daran, die weltweit größte offene Datenbank im Kampf gegen Brustkrebs zu schaffen. Mithilfe der Daten soll eine KI trainiert werden, die hilft, Brustkrebs besser zu erkennen. Mittlerweile hat er anonymisierte Datensätze aus Deutschland, Indien, Afrika, den USA und Südamerika zur Verfügung gestellt bekommen. Die Daten kommen von privaten Kliniken, Unikliniken sowie Krebsforschungszentren, berichtet de Witte. Bis Ende dieses Jahres sollen weitere hinzukommen. Im nächsten Jahr soll dann ein weltweiter Wettbewerb starten, bei dem jeder eigene Analyse-Tools auf den gesamten Datensatz loslassen kann. Dem Open-Source-Gedanken folgend muss jedoch jeder, der Zugriff auf die Informationen haben will, eine Verpflichtung eingehen: Alle Informationen zum KI-Modell, das anhand der Informationen trainiert wurde, müssen veröffentlicht werden. De Wittes Hoffnung: „Wenn wir die attraktivsten Daten zu einer Erkrankung publizieren, dann werden sich alle darauf stürzen."

Auf dem Weg zur Entwicklung einer KI-Lösung, die Mediziner unterstützt, kommt es vor allem auf den Einsatz strukturierter qualitativ hochwertiger Daten an. Das weiß auch Simon Weidert. Er ist Facharzt für Orthopädie und Unfallchirurgie am Klinikum der Universität München (LMU) und Co-Geschäftsführer bei M3i, einer sogenannten Industrie-in-Klinik-Plattform. M3i vernetzt Kliniken und MedTech-Unternehmen. Sie unterstützt Unternehmen während des Entwicklungsprozesses von Produkten und hilft Medizinern, passende Entwicklungspartner zur Umsetzung ihrer Ideen zu finden. Weidert ist daher regelmäßig in Kontakt mit Partnern aus der Industrie und kennt somit auch ihre Herausforderungen beim Thema KI. „Nicht wenige haben sich irgendwo Datensätze besorgt und dann eine blutige Nase geholt", berichtet er. Der Mediziner hat die Erfahrung gemacht, dass viele frei zugängliche Daten nicht zu gebrauchen sind. Der

→

„NICHT WENIGE HABEN SICH IRGENDWO DATENSÄTZE BESORGT UND DANN EINE BLUTIGE NASE GEHOLT … OHNE QUALITÄTSMANAGEMENT KOMMT AM ENDE GAR NICHTS HERAUS."

Dr. med. Simon Weidert, Facharzt für Orthopädie und Unfallchirurgie am Klinikum der Universität München (LMU) und Co-Geschäftsführer bei M3i

Grund dafür sind die Zusatzinformationen, mit denen die Daten versehen sind. Das kann zum Beispiel bei einem CT-Scan die Information sein, wo genau sich ein Tumor befindet. Häufig hätten Mediziner die Zusatzinformationen jedoch nicht einheitlich notiert, kritisiert Weidert. „Ohne Qualitätsmanagement kommt aber am Ende gar nichts heraus", erklärt Weidert. Dann heißt es: Garbage in, Garbage out. Das KI-Modell produziert in diesem Fall lediglich Datenmüll, der niemandem weiterhilft.

M3i gewinnt Daten für seine Projekte in der Regel auf andere Weise. Da die Aufgaben, mit denen Industriepartner auf die Münchener zukommen, häufig sehr speziell sind, gibt es meistens noch keine Daten hierzu. Weidert kontaktiert dann Klinikpartner und klärt mit ihnen, ob sich im Rahmen eines Forschungsprojekts mit einer für sie interessanten klinischen Fragestellung entsprechende Daten sammeln lassen. „Wir versuchen auf diese Weise, sowohl für die Industrie- als auch Forschungsseite einen Gewinn zu erzielen", sagt Weidert. Bevor Daten mit Industriekunden geteilt werden können, muss aber erst noch die zuständige Ethikkommission ihr Okay geben. Hebt die Kommission den Daumen, gehen die Daten zur Anonymisierung in eine Clearingstelle. So wird sichergestellt, dass sich anhand der Informationen nicht herausfinden lässt, um welchen Patienten es sich dabei handelte.

Welche Herausforderungen es bei der Datengewinnung für KI-Projekte gibt, weiß auch Professor Daniel Sonntag. Er leitet den Forschungsbereich Interaktives Maschinelles Lernen am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI). Im Rahmen einer internationalen Kooperation hat das DKFI ein KI-System zur SARS-CoV-2-Diagnose entwickelt. Der Algorithmus wurde dabei mithilfe eines öffentlich verfügbaren Testdatensatzes mit 2.500 CT-Bildern trainiert – und lag später bei seinen Diagnosen sehr häufig richtig: Das KI-System erkannte jeweils mehr als 90 Prozent der Infizierten und Nicht-Infizierten richtig. „Diese Werte sind allerdings mit Vorsicht zu genießen, da der spezielle Testdatensatz, den wir zur Verfügung hatten, nicht repräsentativ ist", sagt Sonntag. Er schätzt, dass ein repräsentativer Datensatz 10- bis 50-mal größer sein müsste – je nachdem, wie unterschiedlich die Bildaufnahmeverfahren in den Kliniken sind. Hilfreich wären künftig harmonisierte Aufnahmeverfahren. Das würde indirekt die Datenqualität verbessern. So ließe sich mit weniger Trainingsdaten auskommen, haben Tests im KI-Labor ergeben, berichtet Sonntag. „Ich denke, diese Tendenz wird sich auch in der Praxis beobachten lassen."

In DFKI-Projekten läuft die Standardisierung zum Beispiel so ab, dass Kliniken oder Unternehmen zu Beginn eines Projekts bereits vorhandene Daten mitbringen. Daraus entsteht eine „Baseline" für den KI-basierten Ansatz. Sprich: Die bisherigen Daten helfen dabei, einen ersten Aufschlag für die Art und Weise festzulegen, wie Daten später gewonnen und erfasst werden. „Innerhalb des Projekts wird dann versucht, über mehrere Datenlieferanten hinweg KI-relevante Standards zu etablieren, die dann innerhalb der Projektlaufzeit und hoffentlich darüber hinaus Einsatz finden", so Sonntag. Wie das in der Praxis abläuft, zeigt sich zum Beispiel in einem DFKI-Projekt zu besseren Diagnose- und Therapieentscheidungen in der Augenheilkunde. Die Kliniken und Technologie-Hersteller haben sich dabei darauf geeinigt, wie die Aufnahmeverfahren der Augenbilder zu entsprechenden Leitlinien innerhalb des Projekts formuliert werden. Dann erhalten alle Mitarbeiter, die die Augenaufnahmen von den Patienten machen, Anweisungen für die zukünftige Datenerzeugung.

Einen Fortschritt für die in Deutschland verfügbaren Daten erhofft sich KI-Experte Daniel Sonntag durch die elektronische Patientenakte (ePA). Denn oft fehlen für neue KI-Lösungen digitale und vollständige Patientendaten. Seit Anfang dieses Jahres können Patienten eine ePA erhalten und verwalten. Sonntag hofft nun, dass sich die ePA in der Breite durchsetze. „Dann könnte man anonymisierte Daten einfacher und rechtskonform erzeugen und für Diagnostik und Versorgungsprozess auswerten."

Damit bei dem Brustkrebs-Diagnose-Projekt von Hippo AI qualitativ hochwertige Daten vorliegen, übernimmt ein Partner die Standardisierung. Die Datenlieferanten schicken lediglich anonymisierte Rohdaten – Pathologiebilder und die dazu relevanten klinischen Befunddaten. Ein Dienstleister aus den USA übernimmt dann das Annotieren. Dabei werden die einzelnen Rohdaten mit Zusatzinformationen versehen – beispielsweise, bei welchen Zellen es sich möglicherweise um Tumorzellen handelt. So soll sichergestellt werden, dass die KIs in dem Projekt mithilfe qualitativ hochwertiger, strukturierter Daten trainiert werden.

Die Brustkrebs-Diagnose soll nun für Hippo AI als Proof of Concept dienen. Gelingt das Projekt, soll das Konzept auch auf andere Erkrankungen übertragen werden. Damit will de Witte dann seiner großen Vision näher kommen: „Wenn wir Open Access ermöglichen, können wir Ungleichheiten weltweit abbauen."

„MIT DER ELEKTRONISCHEN PATIENTENAKTE (EPA) KÖNNTE MAN ANONYMISIERTE DATEN EINFACHER UND RECHTSKONFORM ERZEUGEN UND FÜR DIAGNOSTIK UND VERSORGUNGSPROZESS AUSWERTEN."

Prof. Dr. Daniel Sonntag, Leiter des Forschungsbereichs Interaktives Maschinelles Lernen (IML) am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI)