Kontext

Das FDZ Gesundheit bereitet derzeit die Einführung eines sogenannten Public Use File für Forschende vor. In diesem Zusammenhang wird auch das die Auswertungssoftware im virtuellen Auswertungsraum thematisiert.

Die Arbeitsgruppe (AG) Erhebung und Nutzung von Sekundärdaten (AGENS) der Deutschen Gesellschaft für Sozialmedizin und Prävention (DGSMP) und der Deutschen Gesellschaft für Epidemiologie (DGEpi) sowie die AG Validierung und Linkage von Sekundärdaten des Deutschen Netzwerks Versorgungsforschung (DNVF) haben diese Ankündigung des FDZ zum Anlass genommen, bei ihren Mitgliedern hierzu ein Meinungsbild zu erfragen, mit welchen Softwaretools potentielle Nutzer:innen den obengenannten Public Use File und später im Produktivbetrieb in den virtuellen Analyseräumen des FDZ Gesundheit mit den dort vorhandenen Routinedaten arbeiten würden. Die Umfrage wurde durchgeführt vom 22.07. bis 18.08.2024.

Die Ergebnisse dieser Umfrage werden kommentiert durch den Sprecherkreis der beiden AGs an das FDZ Gesundheit zurückgespielt, um die Prozesse im FDZ Gesundheit zu begleiten und aktiv Einfluss auf eine nutzerorientierte Arbeitsumgebung im FDZ Gesundheit zu nehmen.

Ergebnisse

Die berichteten Ergebnisse basieren auf einer Zwischenauswertung der laufenden Umfrage, Stand 29.07.2024; 13:45 Uhr. Im Ergebnisbericht und für die nachfolgend genannten Zahlen werden die Antworten von N= 169 Teilnehmer:innen gezeigt, die den Fragebogen vollständig ausgefüllt hatten. In Einzelfällen wurde von Befragten berichtet, dass sie beim Ausfüllen der Freitextfelder teilweise eine Timeout-Meldung erhielten und das Ausfüllen des Fragebogens neu starten mussten. Nicht vollständig ausgefüllte Fragebögen (n=71) wurden daher von der Auswertung komplett ausgeschlossen, da sonst bereits registrierte Antworten nach Timeout und Neustart der Abfrage doppelt gezählt worden wären.

Bei der Frage nach der präferierten Software waren Mehrfachnennungen möglich. Kommentare, wie bspw. weitere gewünschte Software die nicht aufgelistet war, konnten in einem Freitextfeld angegeben werden.

Die Ergebnisse der Befragung finden sich in der Anlage als PDF-Dokument. Diese Anlage wurde von Limesurvey standardmäßig erstellt und zeigt die (univariaten) Ergebnisse in Tabellenform und als Grafik. Klartextangaben sind vollständig aufgelistet.

Ergänzend zu den Auswertungen des PDF-Standardreports von Limesurvey finden sich nachfolgend weitere ergänzenden Auszählungen zu den eingesetzten Softwaretools der Nutzer:innen.

Eine Datenbereinigung wurde durchgeführt, indem die Nennung des Tools "SPSS Modeler" auch als SPSS-Nutzer gewertet wurde. Diese Zuordnung erfolgte aus auswertungstechnischen Gründen, da bei den nachfolgend genannten Ergebnissen basierend auf den Auswertungen der vorgegebenen Ankreuz-Antworten ein Fragebogen ohne Toolnutzung geblieben wäre.

R oder Python: Von den 169 Befragten gaben n=124 Personen entweder R oder Python als Softwaretool an; n=7 nur Python, n=77 nur R und n=44 beide Tools.

Statistikpakete MATLAB, SAS, SPSS oder STATA: Von den 169 Befragten gaben n=123 eines oder mehrere der aufgelisteten gängigen Statistikprogramme MATLAB, SAS, SPSS oder STATA an.

Von den n=169 Befragten gaben insgesamt n=167 Personen an, entweder R/Python oder eins der gängigen Statistikprogramm zu nutzen. N=80 gaben an, R oder Python zusammen mit einem der

Statistikpakete anzuwenden, n=44 ausschließlich R oder Python, n=43 ausschließlich mindestens eines der Statistikpakete.

Insgesamt gaben von den n=169 Befragten jeweils die Hälfte an, dass ihnen eine Auswahl ausschließlich von R und Python ausreichen bzw. nicht ausreichen würde.

SQL: Von den n=169 Befragten gaben insgesamt n=56 Personen an, mit SQL arbeiten zu wollen. n=47 nutzen auch der R/Python, n=33 mindestens eines der Statistikpakete. 2 SQL-User gaben an, keine weiteren Tools zu nutzen.

Kommentierung der Ergebnisse aus Sicht der Nutzer:innen

Nahezu jeder Nutzer (167 von 169) gibt an, die FDZ-Daten mit R/Python oder mit einem der gängigen Statistikpakete auswerten zu wollen. Beide Gruppen sind gleich stark vertreten (n=124 vs. n=123). Auch die Gruppengröße der ausschließlichen Nutzer von R/Python oder einem der Statistikpakete ist nahezu identisch (n=44 vs. n =43). Ebenso gibt die Hälfte an, dass ihnen eine ausschließliche Bereitstellung von R und Python nicht ausreichen.

Insgesamt zeigen die Ergebnisse und auch die Vielzahl an geäußerten Kommentaren (n=38), dass potenzielle Nutzer:innen eine Vielfalt an Softwaretools auch jenseits von R nutzen wollen, gerade weil sie damit bereits umfangreiche Erfahrungen mit großen Datenmengen gemacht haben. An erster Stelle nach R wird SAS genannt, etwa jeweils ein Viertel möchte SPSS, STATA oder Python nutzen. Es darf nicht unterschätzt werden, dass bei den komplexen Routinedaten das Erlernen neuer Software ein langwieriger und fehleranfälliger Prozess sein kann, der auch generell nicht empfehlenswert ist (z. B. weil ein Rückgriff auf validierte Standardskripte nicht mehr möglich ist). Auch ob die beiden Tools R und Python mit den großen Datenmengen performant umgehen können, wird angemerkt. Nicht zu unterschätzen ist auch der Aspekt, dass andere Softwaretools einen direkten Einblick in Daten und Variableneinsicht ermöglichen, was nicht nur für Anfänger, sondern auch erfahrene Nutzer:innen extrem hilfreich ist, um sich mit neuen Daten vertraut zu machen.

Insgesamt n=56 von 169 und damit lediglich ein Drittel der Befragten geben SQL als Softwaretool an. Auf den ersten Blick ist dieser niedrige Anteil verwunderlich, liegen die Massendaten des FDZ Gesundheit doch in einer SQL-Umgebung, für deren Abfrage sich SQL als die Standardsprache für die relationale Datenbanken anbietet. Wahrscheinlich (und darauf deuten auch die Kommentare) ist dies vor dem Hintergrund zu interpretieren, dass sowohl R/Python als auch beispielsweise SAS als wichtiger Vertreter der Statistiksoftware die Möglichkeit bieten, SQL-Befehle innerhalb des Tools zu formulieren und an den SQL-Server zu übergeben (Embedding). Somit wurde eine reine SQL-Abfrageumgebung nicht separat als Tool genannt. Voraussetzung für die Nutzung von nativem SQL-Querys in den eingesetzten Tools ist allerdings eine performante Schnittstelle zum SQL-Server. Ist dies der Fall, entfällt das oftmals (lästige) notwendige Wechseln der Analysetools. Wenn mit Nutzung eines einzigen Tools allerdings Performanceabstriche einhergehen, muss überlegt werden, inwieweit die Aufbereitung der Rohdaten mit Zuschnitt eines projektspezifischen Subsamples in einer nativen SQL-Umgebung erfolgt und erst anschließend auf ein Analysetool gewechselt wird. Beides ist bei Implementierung der Softwaretools im FDZ aus Performancesicht zu testen.

Sowohl in R als auch Python stellt die Verwendung von zusätzlichen Paketen ein zentrales Element dar, um den Funktionsumfang der Software stetig zu erweitern. Nennungen von Paketen, welche für den FDZ-Analyseraum gewünscht werden, sind dabei sehr umfangreich und bilden ein erhebliches Funktionsspektrum von generischen Paketen zum Import/Export oder der Aufbereitung von Daten bis hin zu speziellen Anwendungsfällen für Analyse oder Visualisierung von Ergebnissen ab. Es lässt sich daher für die Nutzer eine grundsätzliche Bereitschaft vermuten, weitreichende Arbeitsschritte direkt im FDZ-Analyseraum durchzuführen (z. B. Erstellung publikationsfähiger Tabellen und

Grafiken). Ein Export hinreichend vergröberter Zwischenergebnisse, welche von den Nutzern im Anschluss selbstständig weiterverarbeitet werden, wäre damit von untergeordneter Bedeutung. Gleichsam wird von Nutzern allerdings auch angemerkt, dass im Vorfeld kaum konkret abgeschätzt werden kann, welche Pakete während der Arbeit im FDZ-Analyseraum benötigt werden. Neben der Vielfalt bereits existierender Pakete spielt dabei auch eine Rolle, dass diese ständig fortentwickelt und auch neue Pakete bereitgestellt werden. Es ergibt sich der Eindruck, dass man sich in Hinblick auf Umfang und Flexibilität der Nutzung von Paketen ähnliche Optionen wünscht wie in der lokalen Arbeitsumgebung. Konkret benannt wird für die Arbeit im FDZ-Analyseraum der Zugriff auf ein FDZ-spezifisches Repository oder externes Repository (z. B. CRAN) für den Bezug von Paketen.

Eine Herausforderung im Umgang mit Paketen stellen funktionale Abhängigkeiten dar. So ist beispielsweise in R für die Nutzung bestimmter Pakete oftmals das Vorliegen weiterer Pakete notwendig. Im Rahmen einer Installation von Paketen über die RStudio-Konsole per Zugriff auf ein Repository erfolgt die Mitinstallation weiterer erforderlicher Pakete im Regelfall automatisch, worauf die Nutzer:innen oftmals wenig Einfluss haben und daher die Installation abhängiger Packages nicht explizit benennen (können). Für die manuelle oder automatische Nachinstallation benötigter Packages müssen allerdings geeignete Verfahrenswege, z.B. in Hinblick auf Zuständigkeiten bei Prüfung und Installation von Paketen geschaffen werden. In diesen Prozessen sind Berechtigungen und Verantwortlichkeiten für Download, Installation und Aktualisierung und deren Turnus (auf Wunsch der Nutzer oder regelmäßig zu festgelegten Zeiten) notwendiger Packages zu regeln. Zudem wäre eine systematische Archivierung jeweils genutzter Versionsstände der Statistiksoftware wie auch der Zusatzpakete erforderlich, welche im Bedarfsfall, z.B. Prüfung der Ergebnisse durch Dritte, wiederhergestellt werden müssen. Der hierfür erforderliche Speicherbedarf ist einzuplanen.

Schlussfolgerungen

- Potenzielle Nutzer:innen des FDZ wünschen sich eine Vielfalt an Analysetools. Dabei spielen R und Python eine nennenswerte Rolle, ebenfalls jedoch die klassischen Softwarepakete SAS, STATA und SPSS. Die Hälfte der Befragten beurteilt eine ausschließliche Begrenzung auf R und Python als nicht ausreichend.
- Die für R und Python gewünschte hohe Flexibilität und Personalisierung von Pakten ist eine erhebliche Herausforderung, die vorab zwingend bedacht werden muss.
- Flexible sowie performante Schnittstellen zwischen Tools für Datenmanagement und Analyse sind mitzudenken, insbesondere das performante Zusammenspiel von (nativem) SQL und anderen bereitgestellten Softwareprodukten.
- Unterschiedlicher Anwendungsszenarien bzw. spezifische Tools müssen als transparenter und kontinuierlicher Prozess langfristig in den Betrieb des FDZ Gesundheit implementiert werden.
- Es gibt gute Argumente, den Aufbau der hoffentlich zukunftsweisenden Analyseinfrastruktur beim FDZ mit breitem Einbezug der Nutzer:innen zu gestalten. Die Ergebnisse dieser Umfrage bieten hierfür eine geeignete Diskussionsbasis.

Die AGENS und die AG Datenlinkage bieten den Raum für die zukünftige weitere Konkretisierung des Softwareangebots im virtuellen Analyseraum des FDZ. Auf Veranstaltungen kann bspw. ein Diskussionsforum für zukünftige Nutzer:innen organisiert werden.

Die Sprecher:innen der beteiligten AGs, Enno Swart, Peter Ihle, Holger Gothe, Falk Hoffmann und Stefanie March unter Mitarbeit von Christoph Stallmann und Ludwig Goldhahn. Wir danken insbesondere Christoph Stallmann für die Umsetzung und Administration der Befragung mit Limesurvey.