num compass

Late Mapping Scheme

COMPASS Arbeitspaket 5_Forschungskompatibilität & Interoperabilität

Wenn die ersten Daten erhoben sind, können in aller Regel Datenstrukturen nicht mehr vollständig an einen anderen Datensatz angepasst werden. In solchen Fällen wird das Ausmaß der eventuell vorhandenen GECCO-Konformität der einzelnen Datenelemente geprüft.

Version 1.0_ 16.08.2021

netzwerk universitätsmedizin

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Late Mapping Scheme

Wie wir die Beteiligung von App-Projekten ermöglichen können, die schon vor NUM Compass-Kontakt Daten erhoben haben

Deliverable D5.4B im Rahmen des NUM Compass-Projektes

Autor*innen:
Michael Rusongoza Muzoora
Johannes Oehm
Rasim Atakan Poyraz
Sarah Riepenhausen
Ulrich Sax
Marco Schaarschmidt

Inhaltsverzeichnis

Disclaimer

Ein nachträgliches Mapping sollte vermieden werden, da im Nachgang nie 100% Kompatibilität hergestellt werden kann. Es werden Informationen verloren gehen.

Sollten Sie noch keine Daten erhoben haben, setzen Sie bitte Ihre Fragebögen direkt GECCO-konform um, damit der aufwändige, verlustreiche Prozess, der im Folgenden beschrieben wird, vermieden werden kann. Hinweise dazu finden Sie im “First Contact Package”.

Einleitung

Wenn die ersten Daten erhoben sind, können in aller Regel Datenstrukturen nicht mehr vollständig an einen anderen Datensatz angepasst werden. In solchen Fällen wird das Ausmaß der eventuell vorhandenen GECCO-Konformität der einzelnen Datenelemente geprüft.

Dafür werden die Datenelemente des/der eigenen Fragebögen mit den Datenelementen in GECCO verglichen und in verschiedene Mapping-Kategorien eingeteilt. Da schon geringe Abweichungen (s.a. “Metadaten-modelle” im Anhang) zu einer Inkompatibilität führen können, ist das kein trivialer Prozess. Ziel soll sein, mit so vielen Ihrer Daten zur zentralen Plattform beizutragen, wie möglich. Daher ist zu beachten, dass eine GECCO-Konformität den Datenfluss in Richtung GECCO ermöglichen muss. Das Late Mapping Scheme schließt eine umgekehrten Datenfluss aus, weil dieser u.U. wegen einer notwendigen Transformation Ihrer Daten, nicht mehr möglich ist.

Mehr Informationen zu Interoperabilität, zur Nutzung von simplifier.net und dem Prozess der GECCO-konformen Entwicklung im First Contact Package.

Kategorisierung

Transformierungen finden immer von den bereits erhobenen Daten (Quelle) zu GECCO (Ziel) statt.

Kategorie	**Beschreibung**	**Beispiel**
Identisch	Das Konzept ist identisch zu einem bereits existierenden GECCO-Profil.	Konkrete GECCO-Profile: • Item Lung Disease • Item VerificationStatus
Ähnlich - ohne Transformation	Das Konzept ist ähnlich zu einem bereits existierenden GECCO-Profil, aber die Struktur/Implementierung der Datenerhebung ist unterschiedlich.	Asthma: ja/nein/unbekannt (VerificationStatus) (2 Datenelemente in GECCO, aber in der Quelle ist es nur ein Datenelement Asthma)
Ähnlich - mit Transformation	Das Konzept ist ähnlich zu einem bereits existierenden GECCO-Profil, aber eine Transformation möglich.	• Temperatur in °F vs. °C • Gleitkommazahl in der Quelle, in GECCO Ganzzahl • genaue Werte/Typen für eine Krankheit in erhobenen Daten, passende Klassen in GECCO
Keine direkte Transformation möglich	Das Konzept ist ähnlich zu einem bereits existierenden GECCO-Profil, aber eine Transformation ist nicht direkt möglich. Eine Diskussion mit dem Nachhaltigkeitsausschuss ist notwendig, ob die Daten transformiert werden können oder nicht.	• nicht identische ValueSets ○ männlich/weiblich/divers vs. männlich/weiblich/divers/unbestimmt/k.A. in GECCO → ist divers wirklich das gleiche oder kann es in Fall 1 auch unbestimmt sein? ○ Asthma/COPD/Sonstige vs. Asthma/COPD/Lungenfibrose/ Sonstige → Sonstige eigentlich nicht mappbar, da es in erstem Fall auch Lungenfibrose bedeutet ○ Erhebung Alter in Klassen, aber genaue Werte in GECCO ○ Frage nach Diabetes mellitus (allg.) oder spezifisch nach Diabetes mellitus Typ II
Kein GECCO-Mapping möglich	Das Konzept ist thematisch nicht in GECCO enthalten.	Psychische Auswirkungen der Erkrankung (weil es nicht in GECCO spezifiziert wird)

Direkt auf Variablen aus GECCO mappbar bedeutet den geringsten Aufwand. Das gilt für identische Variablen und sehr ähnliche Variablen, bei denen lediglich die Struktur/Implementierung der Datenerhebung unterschiedlich zu GECCO ist. In beiden Fällen ist keine Transformation notwendig.

Müssen Daten transformiert, also umgerechnet, kategorisiert oder klassifiziert werden, ist der Aufwand größer. Ein Informationsverlust kann – gerade bei Kategorisierungen/Klassifizierungen – nicht immer ausgeschlossen werden, er ist aber möglichst gering zu halten.

Daten, welche keine direkte Transformation erlauben, sind hier am problematischsten. Sie kommen thematisch in GECCO vor, werden aber nicht kompatibel oder auf einer anderen Definitionsebene erhoben. Sie sind damit so unterschiedlich, dass ein Datenfluss nicht möglich wäre, ohne Ungenauigkeiten oder sogar Fehler zu riskieren. Diese Datenelemente erfordern daher eine sehr genaue Prüfung und Diskussion, ob die Daten doch irgendwie in GECCO genutzt werden können.

Das gilt auch für Datenelemente, die (thematisch) nicht in GECCO enthalten sind, weil sie sehr studienspezifisch sind. Ein Beispiel: die psychische Belastung des Patienten durch die Erkrankung wird in GECCO nicht abgefragt. Je nach Fragestellung der Studie sind solche Variablen natürlich absolut sinnvoll. Sie sind aber für den Datenfluss in Richtung GECCO unkritisch. Es lohnt sich allerdings u.U. ein Blick in GECCO PLUS, da dieser Datensatz deutlich größer ist als GECCO selbst.

“Kochrezept”

1. Aufstellung der eigenen Datenelemente/Variablen

Es erfolgt eine Aufstellung der eigenen Datenelemente/Variablen mit folgenden Spalten:

- Frage/Aufforderung/zugrundeliegendes Konzept
- Datentyp (Freitext, Datum, Datum mit Uhrzeit, Zeitdauer, Ganzzahl, Gleitkommazahl, Ankreuzen (Boolean), …, Codelisten/ValueSets mit Einfach-/Mehrfachauswahl)
- Antwortmöglichkeiten (Auswahllisten/ValueSets/Codelisten)
- Einheiten
- optional: Wertebereich
- falls vorhanden: semantische Annotation (z.B. LOINC, SNOMED CT, ICD-10)
- Beispiel/Vorgabe von Format

Hinweis zum Datum:
Jede Ressource benötigt ein Datum in irgendeiner Form. Die Benennung und auch Bedeutung der Datenelemente kann dabei unterschiedlich sein. “recordedDate” ist der tatsächliche Zeitpunkt der Datenerhebung für Conditions, “effective date / time” für Observations. Es kann jedoch auch eine Zeitspanne eingetragen werden, z.B. für Procedures. Eines von beiden, Zeitpunkt oder Zeitspanne, muss angegeben werden. Für Laborwerte und Vitalzeichen entspricht das Datum dem Messzeitpunkt. Wenn beispielsweise eine Krankenschwester über eine Woche verteilt an fünf Tagen Messungen vornimmt und die Daten an einem Tag in die Datenbank eingibt, werden sie als fünf Messungen desselben Tages dargestellt.

2. Vergleich mit GECCO

Nach dem Fertigstellen der Aufstellung, wird diese mit dem GECCO-Datensatz verglichen. Falls FHIR-Erfahrung vorhanden ist, kann ein Abgleich mit den GECCO-Profilen auf simplifier.net (s.a. First Contact Package) erfolgen. Alternativ existiert ein Tabellendokument (Beispieltabelle: https://drive.google.com/file/d/1pbzkpVl8-M6fjRic0nRzr4nMBAp6rCnF/view?usp=sharing) für den Vergleich, wobei keine FHIR-Kenntnis notwendig ist. Dieses ist wie folgt aufgebaut:

- Tabellenblatt “Datenelemente”:
  - Dieses Tabellenblatt enthält die Pflicht-Datenelemente und ihre Annotationen.
  - Falls hier keine Daten geliefert werden können, kann im Ausnahmefall jeweils das DataAbsentReason-Element verwendet werden (Wert z.B. “Antwort unbekannt, weil nicht erfragt”; unter “Optionale Datenelemente”).
- Tabellenblatt “Antwortoptionen”:
  - Dieses Tabellenblatt enthält die Antwortoptionen, die in der gleichnamigen Spalte im Tabellenblatt “Datenelemente” verlinkt sind, mit ihren Annotationen.
- Tabellenblatt “Optionale Datenelemente”:
  - Dieses Tabellenblatt enthält Datenelemente, die zusätzlich in GECCO eingetragen werden können, aber nicht Pflicht sind oder für die ein Systemwert verwendet werden kann, falls keine Daten spezifisch dazu vorhanden sind.
  - Beispiele für ersteres sind viele Datum-Items, das Data-Absent-Reason-Item oder BodySite.
  - Beispiel für das zweite ist das Item mit dem Datum der Datenerhebung (notfalls kann dort das Datum des Datenbankeintrags verwendet werden).

Folgende Fragestellungen gilt es beim Ausfüllen der Tabelle zu beachten:

- Gibt es thematisch etwas passendes zur eigenen Variable?
  - Falls ja, kann die eigene Variable der GECCO-Variable zugeordnet werden.
  - Falls nein, kann die eigene Variable in die Spalten für die “eigenen” Daten (gelb) eingetragen werden.
- Wie wird die thematisch passende GECCO-Variable erhoben?
  - Stimmen Datentyp, Value-Sets, Einheiten und Annotation/Konzept/Inhalt der eigenen Variable mit GECCO überein?
  - Entsprechend werden die o.g. Kategorien für die einzelnen Datenelemente nach Vollständigkeit der Übereinstimmung ausgewählt (Identisch, Ähnlich - ohne Transformation, Ähnlich - mit Transformation, Keine direkte Transformation möglich, Kein GECCO-Mapping möglich).
  - Die Kategorie wird in der Tabellenspalte für die Kategorisierung ergänzt.

Vermeintlich nicht mappbare Items sollten ebenfalls aufgelistet werden.

3. Senden der Tabelle an das Team

Die Tabelle wird an das Nachhaltigkeitsteam gesendet (Kontaktinformationen auf folgender Webseite: https://num-compass.science/de/).

4. Sichtung durch das Team

Die Tabelle wird vom Nachhaltigkeitsteam gesichtet.

5. Gemeinsame Diskussion zu den kritischen Variablen

Es findet eine gemeinsame Diskussion mit dem Nachhaltigkeitsteam statt. Dabei werden unter anderem die folgenden Fragen geklärt:

- Was kann von den Datenelementen, bei denen keine direkte Transformation zu GECCO möglich ist, noch genutzt werden?
- Geht durch die Transformation (zu viel) Information verloren?

Es kommt vermutlich zu Einzelfallentscheidungen bei den Datenelementen, bei denen keine direkte Transformation zu GECCO möglich. Dabei wird ein pragmatisches Vorgehen notwendig (Ist es den Aufwand wert?).

Anhang

Metadatenmodelle

Metadatenmodelle legen fest, in welcher Form die erhobenen Daten am Ende vorliegen sollen. Sie sind eine Zusammenstellung von allen Fragen und Aufforderungen, den Variablen bzw. Datenelementen, die zu dem Thema des Formulars/Modells erhoben werden sollen. Es sind quasi die leeren Formulare oder Datenerhebungsbögen einer Studie, eines Registers oder einer Station in einem Krankenhaus.

Jede Variable wird dabei mit Datentyp (Freitext, Ganzzahlen, Gleitkommazahlen, Datum, …), Antwortoptionen, erlaubten Werte(bereiche)n, Einheiten und der formulierten Fragen/Aufforderungen definiert. Je nach Datenformat kann diese Formulierung auch auf mehr als einer Sprache vorliegen. Um die Bedeutung der Daten maschinenlesbar und sprachunabhängig zu hinterlegen, können semantische Annotationen hinzugefügt werden. Dies können z.B. Codes aus den medizinischen Terminologien SNOMED CT und LOINC, aber auch ICD-10-Codes sein.

Ziel - und Herausforderung - sind dabei kompatible Metadatenmodelle, damit die Daten nach der Erhebung gemeinsam ausgewertet werden können. Dadurch werden studienübergreifende Analysen einfach/direkt möglich, die Daten sind also interoperabel.

Optimalerweise sollte diese Kompatibilität der Datensätze schon während der Entwicklung der Metadatenmodelle berücksichtigt werden. Wo möglich sollte dafür auf Standards zurückgegriffen werden. Im Fall einer Pandemie bietet sich hier der GECCO-(PLUS-)Datensatz an. Elemente, die auch in GECCO vorkommen, sollten so wie in GECCO erhoben und GECCO-Elemente, die ergänzend zur eigenen Erhebung passen würden, zusätzlich übernommen werden. Studienspezifische Elemente, die nicht in GECCO enthalten sind, sollten sich an anderen nationalen oder internationalen Standards orientieren, wenn möglich, oder zumindest wie in anderen Studien erhoben werden.

Damit die eigenen Metadaten verfügbar sind um als Vorlage für andere zu dienen, sollten sie veröffentlicht werden. Dafür bieten sich öffentliche Plattformen an, wie z.B. simplifier.net für FHIR-Profile oder das Portal für medizinische Datenmodelle (MDM-Portal, https://medical-data-models.org) für ODM-Dateien/zunächst nicht als FHIR-Profile angelegte Darstellungen der Metadatenmodelle. Das ermöglicht dann dort die Recherche nach der Umsetzung von anderen und die direkte kompatible Erhebung. Vorteil einer zentralen Veröffentlichung ist, dass nicht alle Projektseiten, die evtl. vorhanden sind, einzeln überprüft oder die Verantwortlichen kontaktiert werden müssen.

Das Problem bei fehlender Kompatibilität ist, dass die schon erhobenen Daten gemappt werden müssen. Das führt u.U. schon bei geringen Abweichungen zu Informationsverlust. Ein einfaches Beispiel dafür ist die Erhebung eines Symptoms. Studie A nutzt dafür eine Skala von A bis D, Studie B eine Skala von 0 bis 4.

Studie A: Symptom X	A		B		C		D
Studie B: Symptom X	0	1		2		3	4

Die Minimal- und Maximalwerte können evtl. noch gemappt und zusammen ausgewertet werden, wenn hinterlegt ist, dass A bzw. 0 das Minimum und D bzw. 4 das Maximum des Symptoms ist. Die Werte dazwischen sind aber nicht mappbar, da ihre Bedeutung zu unterschiedlich ist, und eine gemeinsame Auswertung ist kaum/nicht möglich.

Folgende Universitätskliniken des Netzwerks Universitätsmedizin nehmen am COMPASS-Projekt teil:

Charité – Universitätsmedizin Berlin
Universitätsmedizin Göttingen
Universitätsmedizin Mainz
Universitätsklinikum Würzburg
Uniklinik Köln
Universitätsklinikum Münster
Universitätsklinikum Regensburg
Universitätsklinikum Ulm
Universitätsklinikum Erlangen

Ansprechpartner für weitere Fragen:
COMPASS Koordinierungsstelle
compass@unimedizin-mainz.de

num compass

https://num-compass.science

@CompassNum

netzwerk universitätsmedizin

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung