DEMIS – Ressortforschungsantrag – Signale 2.0

Erweiterung des automatischen Früherkennungssystems zu einem Ausbruchsinformations-System durch datengetriebene Kontextgewinnung mit Hilfe aktueller Methoden des Maschinellen Lernens und Integration in DEMIS

Dr. Hermann Claus, Dr. Göran Kirchner,
Dr. Alexander Ullrich, Dr. Stéphane Ghozzi
Robert Koch-Institut

Abteilung für Infektionsepidemiologie
Fachgebiet Datenmanagement (FG 31)

Oktober 2017

Das aus dem Forschungs-Projekt Signale (1.0) hervorgegangene Frühwarnsystem wird routinemäßig von EpidemiologInnen des RKI und anderen Mitarbeitenden des ÖGD genutzt, um mögliche Ausbrüche zu identifizieren. Ziel dieses Forschungs-Projektes ist es, das Frühwarnsystem in DEMIS zu integrieren und so zu erweitern, dass ein komplettes, mehrdimensionales Bild einer epidemiologischen Lage erstellt werden kann. Dadurch sollen EpidemiologInnen darin unterstützt werden, schnell die richtigen Entscheidungen zu treffen. Dies ist besonders in Krisen-Situationen wichtig. Gleichzeitig sollen die Dienste auch im Rahmen von DEMIS anderen Akteuren (insbes. Meldenden) als Mehrwertdienst zur Verfügung gestellt werden, um so die Attraktivität zu erhöhen und die Compliance zu verbessern.

Inhaltsverzeichnis

1 Zusammenfassung

Das aus dem Signale-Projekt (Kapitel 1501 Titel 54401, 11/2015 – 12/2017) hervorgegangene Frühwarnsystem wird routinemäßig von EpidemiologInnen des RKI und anderen Mitarbeitenden des ÖGD genutzt, um mögliche Ausbrüche zu identifizieren. Ziel dieses Projekts ist es, das bereits erfolgreich implementierte Frühwarnsystem um weitere Funktionalitäten zu erweitern und in das *Deutschen Elektronischen Melde- und Informationssystem für den Infektionsschutz* DEMIS zu integrieren, sodass ein komplettes, mehrdimensionales Bild der epidemiologischen Lage erstellt werden kann. Dadurch sollen EpidemiologInnen darin unterstützt werden, Ereignisse besser bewerten und insbesondere in Krisensituationen schneller handeln zu können.

Das System soll von einer reinen statistischen Ausbruchserkennung in ein Werkzeug zur umfassenden *Ausbruchsinformation* übergehen. Dafür werden die Signalinformationen mit relevanten Kontextinformationen angereichert. Das bedeutet, dass nicht nur wie bisher die IfSG-Falldaten (Anzahl von Fällen in einer bestimmten Region und Bevölkerungsgruppe) berücksichtigt werden, sondern auch eine Vielzahl verschiedenartiger ergänzender Datenquellen. Dies umfasst relevante Datenquellen von strukturierten Daten (z.B. meteorologische Daten, syndromische Surveillance) über kurierte unstrukturierte Daten (z.B. Fachpublikationen, intern und extern) bis hin zu Social-Media-Daten (z.B. Suchanfrage, Sentiment-Analysen). Dafür werden moderne Methoden des maschinellen Lernens und des Natural Language Processings (NLP, dt. Computerlinguistik) verwendet und neue statistische Methoden entwickelt (Anwendungsforschung). Zur effizienten Nutzung der gewonnenen Informationen wird eine interaktive und intuitive sowie personalisierte Oberfläche zur Verfügung gestellt.

Das Projekt soll als *wesentlicher Mehrwertdienst* in die gegenwärtig entstehende DEMIS-Landschaft integriert werden. Die Diversität der Nutzenden des DEMIS-Systems und ihrer Anforderungen wird in allen Projektabschnitten berücksichtigt. Damit ist es eine wichtige Ergänzung und nötige Erweiterung der im Rahmen von DEMIS entstehenden Infrastruktur und trägt entscheidend zur Attraktivität des Gesamtsystems bei.

Durch das Forschungs-Projekt werden die vorhandenen Kompetenzen des Robert Koch-Instituts (RKI) im Bereich *Data Science* (Daten-Bereitstellung, maschinellen Lernens, anspruchsvoller Visualisierungen) genutzt, weiter ausgebaut und kommen perspektivisch allen Akteuren innerhalb des Meldesystems zu Gute. Auf dem Gebiet des *maschinellen Lernens* und der *Compuerlinguistik* sollen mathematische Verfahren erforscht und weiterentwickelt werden, die das Thema der *Digitalen Epidemiologie* der *Strategie 2025 des RKI* deutlich voranbringen. Die Werkzeuge und Dienste werden mit besonderer Sorgfalt öffentlich bereitgestellt, nach einer verantwortungsvollen Open-Data- und Open-Source-Philosophie, die die gesetzlichen Rahmenbedingungen beachtet. Es trägt damit auch zur Etablierung von zeitgemäßen Formen des *Forschungsdatenmanagements* bei.

2 Ziele des Projektes

Ziel ist es den bereits existierenden Dienst zur automatischen Früherkennung von Ausbrüchen [21] zu einem System zur *Ausbruchsinformation* zu erweitern und allen NutzerInnen von DEMIS zur Verfügung zu stellen. Das System soll eine *interaktive Echtzeitanalyse* aller relevanten Daten ermöglichen. Dabei sollen die verschiedenen Anforderungen der unterschiedlichen Nutzerrollen (RKI, Landestellen, Gesundheitsämter, Labore, Ärzte und anderen meldepflichtigen Einrichtungen) berücksichtigt werden.

Die IfSG-Meldedaten und Signalinformationen sollen desweiteren mit sinnvollen *Kontextinformationen* angereichert werden. Dafür sollen relevante interne und externe Datenquellen identifiziert, extrahiert und für eine anschließende Analyse aufbereitet werden. Für die Verarbeitung unstrukturierter Daten sollen aktuelle Methoden des *Natural Language Processings* verwendet werden. Die aufbereiteten Daten sollen auch anderen DEMIS-Anwendungen zur Verfügung gestellt werden.

Die bisherigen *Signalerkennungs-Algorithmen* sollen die so bereitgestellten Daten integrieren. Weiterhin werden neue adaptive Vorhersagealgorithmen entwickelt, die eine *Vorhersage* und den Vergleich mit Daten aus der *Simulation von hypothetischen Szenarien* erlauben. Dafür werden jeweils verschiedene Ansätze des *maschinellen Lernens* getestet werden.

Das neue System soll vollständig in die Infrastruktur von *DEMIS integriert* werden, um die dort vorhandenen umfangreichen Daten und Funktionalitäten zu nutzen, sowie die Signalinformationen und die integrierten Datenquellen weiteren DEMIS-Anwendungen und deren Anwendern zur Verfügung zu stellen. Insbesondere die Dienste, die im Rahmen des *DEMIS-Semantik-Projekts* entstehen, sollen verstärkt genutzt werden. Diese semantischen Technologien zusammen mit modernen Verfahren der Computerlinguistik sollen es ermöglichen, einen *komfortablen Zugang* zu den bereitgestellten Informationen zu bieten.

Im Rahmen des Projektes soll auch die Nutzung von etablierten Plattformen zur Verarbeitung natürlichsprachlicher Daten, wie etwa *IBM Watson Analytics* oder *Natural Language Toolkit* (NLTK, [1]), erschlossen werden. Die Hauptgründe für den Einsatz solcher Plattformen sind in der Verknüpfung von den gut strukturierten und am RKI vorhandenen Daten mit unstrukturierten oder nicht maschinell-verarbeitbaren Datenbeständen durch computerlinguistische Methoden zu finden. Ein deutlicher Mehrwert ist mit einem natürlichsprachlichen Zugang zu den RKI Angeboten zu erwarten. Die niedrigschwelligen Möglichkeiten zur Erzeugung von interaktiven Berichten sollten ausgenutzt werden. Weniger die in *IBM Watson Analytics* zur Verfügung gestellten Daten aus sozialen Netzen, als die generelle Möglichkeit, unstrukturierte Daten durch Lernverfahren der Künstlichen Intelligenz zu erschließen, bieten ein enormes Potenzial.

Durch den Einsatz von *Empfehlungssystemen (recommender systems)* und anderen Komponenten der *Personalisierung* sollen die aufbereiteten Informationen nach den individuellen Bedürfnissen der Nutzenden von DEMIS zur Verfügung gestellt werden. Gleichzeitig wird hierdurch auch die Möglichkeit geschaffen, das Feedback der verschiedenen Nutzergruppen einzubinden und zeitnah (real-time) zu verarbeiten.

2.1 Ergebnisse (Deliverables)

Die Ergebnisse sollen in internationalen Fachzeitschriften unter Federführung des wissenschaftlichen Projektleiters des RKI veröffentlicht werden. Darüber hinaus sollten Teile als Open-Source-Software veröffentlicht werden.

Am Ende der Projektzeit sind folgende Ergebnisse vorzuweisen:

- Es wurden neue praxisrelevante Methoden entwickelt und publiziert bzw. eingereicht.
  - mind. 2 Publikationen in Fachzeitschriften oder Journals
  - mind. 2 open-source Methoden-Veröffentlichungen (z.B. als R-Paket auf CRAN bzw. gitlab.com)
  - mind. 2 REST-API als Beitrag zur Open-Data-Initiative
- Das DEMIS Informationssystem ist durch die entwickelten Methoden bereichert (RKI, Landestellen, Gesundheitsämter, Labore, Ärzte und anderen meldepflichtigen Einrichtungen).
  - Kontextinformationen (mind. 2 aus dem Bereich strukturierter Daten, mind. 1 aus dem Bereich unstrukturierter Daten)
  - Aussagekräftige Bewertungsverfahren (Scores, inkl. Evaluation)
  - Personalisierte Informationsbereitstellung (inkl. Evaluation)
- Ein Dashboard ermöglicht eine *komfortable und personalisierte* Arbeit mit Meldungen, Fällen, Ausbrüchen und Signalen, die um wertvolle Kontextinformationen angereichert sind.
  - informative und interaktive Visualisierungen (inkl. Evaluation)

Weitere Details zu den Deliverables sind in der Beschreibung der Arbeitspakete zu finden.

3 Problemhintergrund

3.1 Problemhintergrund und Wissensstand

Das Robert Koch-Institut hat bisher mit SurvNet@RKI bereits ein elektronisches Verfahren zur Eingabe, Übermittlung und Analyse der Daten zu übertragbaren Krankheiten, die gemäß Infektionsschutzgesetz (IfSG) meldepflichtig sind, entwickelt. Im März 2013 führte die Änderung des IfSG zu einer erheblichen Verkürzung der Übermittlungsfristen für Gesundheitsämtern und Landestellen. Diese Änderung des Gesetzes war auch eine Antwort auf den STEC-O104:H4-Ausbruch in 2011, infolge dessen die Dringlichkeit einer frühzeitigen Erkennung von Ausbrüchen deutlich wurde.

Mit dem *Deutschen Elektronischen Melde- und Informationssystem für den Infektionsschutz* wird das existierende Meldesystem für Infektionskrankheiten gemäß IfSG weiterentwickelt und verbessert. Insbesondere wird – beginnend bei den Meldenden (Ärztinnen und Ärzte, Labore, andere) – eine durchgängig elektronische Informationsverarbeitung ermöglicht. Dadurch soll der Aufwand für die Meldenden und die zuständigen Behörden reduziert werden und Informationen zu auftretenden Infektionskrankheiten können künftig schneller bei den Verantwortlichen in den Gesundheitsämtern, den zuständigen Landesbehörden und am RKI vorliegen. Weiterhin werden die Zusammenarbeit der Beteiligten und der Datenaustausch zwischen ihnen besser unterstützt, sodass auch große Infektionsereignisse effektiver bearbeitet werden können.

In den vergangenen Jahren wurde die *automatisierte Analyse* der IfSG-Falldaten ausgebaut. Hierzu wurde ein erstes System [21] entwickelt, welches ca. 80% aller übermittelten Fälle jeden Tag mit dem Ziel auswertet, Ausbrüche und andere unerwartete Häufungen frühzeitig zu erkennen. Die Erfahrung hat gezeigt, dass die Masse an Fällen nicht allein durch manuelle Verfahren adäquat analysiert und überwacht werden kann. Das aktuelle System hat die Möglichkeiten verdeutlicht, die ein solches System für den Öffentlichen Gesundheitsdienst bietet. Trotzdem existiert noch ein großes Potential für den nachhaltigen Ausbau eines solchen Systems am RKI und im Rahmen von DEMIS.

In welchem Maße eine mögliche DEMIS-Infrastruktur vom Einsatz moderner und praxisrelevanter Verfahren profitieren kann, wurde bei der bisherigen Umsetzung aufgezeigt. Die Signalgenerierung am RKI bildete dabei die Grundlage für die benutzerspezifische signalbasierte Benachrichtigung über Auffälligkeiten bei Meldungen. Bei der Umsetzung von DEMIS wird dieser Baustein eine entscheidende Rolle spielen, da hier das Informationsaufkommen noch einmal höher sein wird (ca. 2000 Arzt- und Labor-Meldungen täglich, im Vergleich zu 1000 übermittelten Fällen bisher) und die Daten auch mit einer deutlich geringeren zeitlichen Verzögerung zur Verfügung stehen.

Das Thema der *Digitalen Epidemiologie* ist einer der Grundpfeiler der Agenda 2025 des Robert Koch-Instituts und der damit verbundenen Bestrebungen. Mit den gegenwärtigen Trends in *Big Data*, *Social Media* u.ä. werden neue mathematische Verfahren entwickelt, die die klassischen statistischen Verfahren ergänzen, um die zeitnahe Analyse großer Mengen oder unstrukturierter Daten zu ermöglichen. Insbesondere für die Epidemiologie ergeben sich hieraus vielversprechende Perspektiven [2, 19]. Diese Verfahren sollen im Rahmen des DEMIS-Informationsdienstes verfügbar gemacht, weiterentwickelt und einem größeren Nutzerkreis zugänglich gemacht werden. Die erzielten Forschungsergebnisse sind neben dem praktischen Einsatz auch von wissenschaftlichem Wert.

Durch eine enge Kooperation mit universitären Forschungsgruppen soll ein solider und qualitativ hochwertiger Beitrag des RKIs in den genannten Forschungsgebieten geleistet werden.

3.2 Genderaspekte

Es sind keine geschlechtsspezifischen Aspekte für das Projekt von Bedeutung.

3.3 Vorarbeiten und Vorleistungen

Im Verlauf des derzeitigen Signale-Projekts (11/2015 - 12/2017) wurde ein *Früherkennungssystem* etabliert, das automatisch und tagesaktuell Auffälligkeiten in Zeitreihen von mehreren hunderttausenden Filterkombinationen (Suchkriterien nach Erreger, Subtyp, Region, Altersgruppe und Geschlecht) entdeckt, speichert, aufbereitet und zusammenfassende Berichte an Mitarbeitende des RKI und andere Mitarbeitende im ÖGD versendet. Das System und die Ergebnisse sind in zwei Publikationen dokumentiert [20, 21].

Das System wurde modular umgestaltet, um eine hohe Robustheit und Flexibilität zu gewährleisten. Somit können neue Funktionalitäten hinzugefügt werden, ohne den Betrieb des Signaldienstes einzustellen. Es wurden Algorithmen entwickelt, die die manuelle Ausbruchszuordnung einbeziehen und die Integration von externen Datenquellen erlauben. Erste Ergebnisse wurden dokumentiert in [8] und auf dem SIREMTI Workshop präsentiert. Am RKI wurde im September 2017 der internationale Workschop „Automatic Detection of Infectious Disease Outbreaks“ durchgeführt. Hier wurden die eigenen Entwicklungen vorgestellt, Kooperationen initiiert und zukünftige Vorhaben auf diesem Gebiet diskutiert (s.a. Anlagen *Workshopbericht*, *Letters of Intent*). Die Ergebnisse werden auf der ESCAIDE 2017 präsentiert. Zwei weitere Veröffentlichungen sind in Arbeit.

Weiterhin wurden im Bereich der Zeitreihenanalyse Erfahrungen mit gängigen statistischen Verfahren und modernen Methoden des maschinellen Lernens gesammelt. Daraus ergeben sich wertvolle Erkenntnisse bezüglich der Wahl von Werkzeugen, Algorithmen und Methoden, sowie der angemessenen Umsetzung der Nutzeranforderungen. Ein Prototyp eines interaktiven Live-Dashboards zur Darstellung von Signalinformation und IfSG-Meldedaten wurde entwickelt und wird derzeit RKI-intern getestet. Erste Methoden zur Etablierung eines Empfelungssystems wurden im Rahmen einer Bachelor-Arbeit ausgearbeitet [9].

Für den Dashboard-Prototyp wurde eine Schnittstelle (API) zum SurvNet-Cube (OLAP), einem Auswerte-Tool für die Meldedaten, genutzt. Diese Technologie bietet eine Möglichkeit der Signalberechnung auf effiziente Weise, die bisher noch nicht genutzt wird, aber zur Etablierung eines Echtzeitsystems notwendig ist. Die Software SurvNet@RKI [6, 12] bietet bereits eine rudimentäre Funktionalität der Darstellung von Signalinformationen. Ein Konzept zur Integration der Signalinformationen in den SurvNet-Cube wird derzeit entwickelt. Damit wurde die Integration der Signalinformation in die DEMIS-Struktur vorbereitet. Für den produktiven Einsatz ist hier jedoch noch weitere Arbeit zu leisten.

Zusätzlich zum bisherigen Signale-Projekt verfügt das beantragende Fachgebiet über Kenntnisse und Erfahrungen aus den benötigten Bereichen, wie dem Machinellen Lernen und der Computerlin-

gustik. So beteiligte es sich am EU-Projekt M-Eco (Medical Eco-System) [4], in dem unstrukturierte Daten aus Web-2.0-Inhalten (etwa twitter.com) analysiert und in die Bewertung von Ausbrüchen einbezogen wurden. Methoden, in denen Algorithmen des Machinellen Lernens eingesetzt werden, sind ebenso bereits im produktiven Einsatz. So entstanden Verfahren zur automatisierte Zusammenführung von anonymen Meldedaten, die gemäß §7.3 IfSG an das RKI gemeldet werden.

Nicht zuletzt verfügt das RKI mit der Entwicklung und dem Vertrieb des Gesundheitsamt-Fachverfahrens SurvNet@RKI seit 2001, dem System zur Erhebung von Antibiotika-Resistenz-Ergebnissen ARS seit 2007 und dem DEMIS-Pilot-Projekt 2012-2014 wertvolle Erfahrungen im Umgang mit komplexen Systemen und großen Projekten. Der Bereich der Expertise erstreckt sich dabei über verschiedenste Bereiche der Mathematik, Statistik und Software-Entwicklung.

4 Design und methodische Vorgehensweise

Das Arbeitsprogramm gliedert sich in drei Arbeitsbereiche:

1. Entwicklung eines Ausbruchsinformations-Systems
2. Methoden-Entwicklung (Maschinelles Lernen, Statistik, Recommender System, Computerlinguistik/NLP)
3. Vernetzung mit Behörden und Forschungsgruppen im Bereich Public-Health

4.1 Entwicklung eines Ausbruchsinformations-Systems

Automatische Kontextgenerierung

Die Signalinformationen, so wie sie derzeit erzeugt und präsentiert werden, geben zunächst nur Hinweise auf *statistisch auffällige Häufungen* in einer bestimmten Region, für eine gewisse Altersgruppe zu einem gegebenen Zeitraum. Es bleibt die arbeitsintensive Aufgabe der EpidemiologInnen, diese Signale untereinander zu vergleichen und zu bewerten.

Die Anreicherung der Fall- und Signalinformationen durch Kontextinformationen soll die Aufgabe der *Bewertung, Entscheidungsfindung und Maßnahmenergreifung* erleichtern. Relevante kontextuelle Informationen sind zum Beispiel Beziehungen zwischen Fällen durch ähnliche Ausprägungen gewisser Merkmale aber auch externe Faktoren wie das Wetter oder Daten der syndromischen bzw. molekularen Surveillance. Hintergrundinformationen, wie die Krankheitslast (burden of disease) oder extrahierte Informationen aus der Fachliteratur, tragen oft wesentlich zur Entscheidungsfindung bei, sind jedoch selten in einem wünschenswerten Maße verfügbar und ihr Beitrag kann auch im Nachgang nicht systematisch ausgewertet werden.

Diese Datenquellen sollen zukünftig mit den Fall- und Signalinformationen verbunden werden und zu aussagekräftigen Zusammenfassungen genutzt werden. Mit ihnen kann auch die Grundlage zu einer Vorhersage von Krankheitsentwicklungen geschaffen werden.

Die Intuition und Erfahrung der EpidemiologInnen ist immer noch unerlässlich für den Prozess des Ausbruchsmanagements. Die Verfahren zur Aufbereitung und Präsentation der kontextuellen Daten sollen deshalb aus dem Verhalten und der Interaktion der NutzerInnen lernen. Dies beinhaltet sowohl die Auswertung des Nutzerverhaltens als auch explizite Nutzereingaben. Solche Algorithmen existieren bereits sowohl für Business- als auch Public-Health-Anwendungen [7, 11, 24].

Personalisiertes Dashboard und Berichterstattung

Das als Prototyp existierende *Signale-Dashboard* soll um Funktionalitäten erweitert werden, die eine bessere Bewertung der epidemiologischen Situation ermöglichen.

Dabei sollen die Anforderungen der verschiedenen Rollen (RKI, Landestellen, Gesundheitsämter, Labore, Ärzte und anderen meldepflichtigen Einrichtungen) durch eine entsprechende Zusammen-

stellung der für die jeweilige Rolle wichtigen Informationen berücksichtigt werden. Weiterhin soll das Dashboard *personalisierbar* sein, um auch den individuellen Erfordernissen der NutzerInnen innerhalb der gleichen Rolle zu entsprechen.

Die oben erwähnten kontextbezogenen Informationen sollen als situations-angepasster News-Feed (z.B. Zusammenfassung von relevanten Informationen aus Social-Media und Nachrichten) und Bibliographie (Verweise auf Publikationen und internen Dokumenten zu vergleichbaren Public-Health-Ereignissen) bereitgestellt werden. Zusätzlich sollen auch kompakte Darstellungen verschiedener Indikatoren (Social-Media-Index, Sentiment-Index) entworfen werden, die dem Ziel der schnellen Übersichtsgewinnung dienen.

Neben der Darstellungsform soll auch die Art der Interaktion auf unterschiedlichen Wegen ermöglicht werden. Neben dem Dashboard, das versucht alle Informationen kompakt und zusammen darzustellen, soll auch eine sequentielle Darstellung in Form eines Story-Boards, Notebook oder Guides zur Verfügung gestellt werden. So kann der Nutzende nach und nach *von einer groben Übersicht hin zu einer detaillierten Lagebeschreibung* navigieren und dadurch gut-informiert über konkrete Maßnahmen entscheiden. Außerdem soll es möglich sein, die so identifizierten relevanten Informationen und Visualisierungen in Form von Berichten zu exportieren, um die Grundlage für die Entscheidungen angemessen zu dokumentieren.

4.2 Methoden-Entwicklung

Maschinelles Lernen

Die Integration der verschiedenen Datenquellen erfordert die Entwicklung und Anwendung neuer Methoden. Im Bereich des *Maschinellen Lernens* existieren moderne Methoden, die sich besonders für den Umgang mit vielfältigen Datenquellen [26] und großen Datenmengen [25] eignen. Hierzu müssen *existierende* Methoden des Maschinellen Lernens adaptiert, implementiert und an die Besonderheiten der Meldedaten angepasst werden. Dies ist gegenwärtig möglich, da viele der state-of-the-art Verfahren bereits durch führende Technologieunternehmen (z.B. Google und Facebook) implementiert sind und als open-source veröffentlicht wurden. Ebenso soll der bestehende Kontakt mit akademischen Forschungsgruppen, die bereits über relevante Erfahrungen verfügen, weiter ausgebaut werden.

Auch eine *Kurzzeit-Prognose* (Forecast) zusammen mit der damit verbundenen Ungewissheit, würde die Planung epidemiologischer Maßnahmen erheblich unterstützen, insbesondere in Krisen-Situationen. Hierbei spielt die zügige Verfügbarkeit von Daten eine wichtige Rolle. Es gibt gut dokumentierte und frei verfügbare Ansätze die Social-Media-Daten berücksichtigen [3], externe Datenquellen integrieren und Feedback von Anwendern bzw. Experten nutzen [7], um möglichst gute Vorhersagen zu erstellen.

Es soll auf Basis bestehender Ansätze ein lernendes Modell entwickelt werden, das eine Vorhersage der Meldedaten für wenige Wochen anbietet und Nutzereingaben sowie weitere Faktoren berücksichtigt.

Basierend auf den Ergebnissen der Grenzwertberechnung und Kurzzeit-Prognose sowie des weite-

ren Kontextes entscheiden EpidemiologInnen, ob eine Situation eine Gefahr für die Bevölkerung darstellt. Ein computergestütztes System kann aus diesen Beispielen („labeled data“) lernen, was ein Ausbruch ist. Das RKI verfügt über einen weltweit einzigartigen Datensatz: Ausbrüche, als Sammlung von Fällen, werden von EpidemiologInnen in SurvNet seit In-Kraft-Treten des IfSG im Jahr 2001 erfasst [6, 12].

Es soll ein maschinelles Lernverfahren trainiert werden, das mit Hilfe der historischen Ausbruchsdatensätze epidemiologische Einschätzungen anbieten kann.

Auch aus den Interaktionen der EpidemiologInnen mit dem System sollen Informationen für Andere gewonnen werden. Ein Empfehlungsdienst soll dadurch zur Bewältigung der Informationsüberflutung beitragen, indem es aus einer unübersichtlichen Menge an Informationen die wichtigen heraus sucht, um diese dann anderen Nutzenden bereitzustellen.

Es soll ein Recommender-System zur Bewältigung der Informationsüberflutung bereitgestellt werden.

Statistische Methodik

Die Methoden der statistischen Analyse, die für die Signalerkennung verwendet werden, sollen weiterentwickelt und verbessert werden. Die Schwerpunkte sind dabei die *Berücksichtigung des Meldeverzugs* und der raumzeitlichen Verbreitung von Ausbrüchen. Perspektivisch sollen auch Ergebnisse der *molekularen Surveillance* mit einbezogen werden. Die im Vorgängerprojekt identifizierten Ansätze [15, 22] sollen an die Eigenheiten der Meldedaten angepasst werden (z.B. Behandlung von fehlenden Werten) bzw. in neue effizientere Prozesse der Signalgenerierung eingepasst werden (Clustersuche statt vordefinierte Filterkombinationen [13]).

Die Vorarbeiten zur systematischen Bewertung und Optimierung von Signalerkennungs-Algorithmen sollen weiter auf die genannten neuen und angepassten Methoden angewendet werden und um weitere Verfahren ergänzt werden.

Computerlinguistik

Es gibt viele Informationsquellen, die ihre Daten in *unstrukturierter Form* oder zumindest in *maschinell schwierig verarbeitbarer Form* bereitstellen und oft nur als Text in natürlicher Sprache vorliegen. Für das Management von Ausbrüchen relevante Quellen sind zum Beispiel RKI-intern erstellte Berichte wie das Infektionsepidemiologische Jahrbuch und das Epidemiologische Bulletin sowie Mitschriften der EpiLag, Informationen des Meldesystems aus Freitexten in SurvNet@RKI, Warnmeldungen von anderen Institutionen aus ProMED-mail oder den RASFF Meldungen zu Lebensmitteln ebenso wie allgemeine Fachpublikationen zu Erregern und Ausbrüchen.

Aus diesen Quellen sollen Informationen automatisch extrahiert und aufbereitet werden, um sie in die DEMIS-Infrastruktur zu integrieren und für das Ausbruchsinformations-System nutzbar zu machen.

Weiterhin sollen Informationen aus *sozialen Netzwerken* analysiert werden, um die Aktivität und Stimmungslage der Bevölkerung bezüglich bestimmter Public-Health-Ereignisse (Grippewelle, Aus-

brüche, Trends, Migration) zu messen. Dies kann wertvolle Hinweise darauf geben, wann und wo ein besonderer Bedarf an Aufklärung und Kommunikation besteht. Hierfür werden bestehende Schnittstellen [5] genutzt, um relevante Daten zu sammeln, vorhandene Methoden zur Analyse [1, 14] angewandt und sinnvolle Indikatoren entwickelt. Die Erfahrungen aus vergangenen Projekten (M-Eco) soll dabei berücksichtigt werden [4].

Es existieren Ansätze des *Natural Language Processings*, um aus einzelnen Dokumenten oder gar einem ganzen Text-Korpus Informationen zu extrahieren und dafür zu nutzen, um kurze Zusammenfassungen zu generieren, Ähnlichkeiten zwischen Dokumenten oder Konzepten herzustellen und Dokumente zu klassifizieren. Viele dieser Ansätze sind bereits in Softwarebibliotheken frei verfügbar [23] und können in das System integriert werden. Selbst die aktuell besten Methoden stehen teilweise frei zur Verfügung [17, 18] oder sind über eine *Kooperation mit IBM Watson* [10, 16] zu integrieren.

Diese Ansätze sollen prototypisch getestet und nach technischen und wirtschaftlichen Kriterien bewertet werden, um so den erfolgversprechendsten Ansatz auszuwählen und in den produktiven Einsatz zu überführen.

4.3 Vernetzung

Für den Erfolg des Projektes ist ein intensiver Austausch mit anderen Public-Health-Instituten wichtig, um verschiedene Themen zu besprechen und gemeinsam voranzubringen. Die Schwerpunkte sollen dabei auf den Themengebieten:

- Etablierung von Standards für die Bereitstellung von Daten (Formate, APIs und Lizenzen)
- Verwendung von epidemiologischen Diensten und Werkzeugen (z.B. die hier beschriebenen Signalerkennungsdienste)

liegen.

Es wird angestrebt die Vernetzung und Zusammenarbeit sowohl RKI-intern als auch mit anderen Public-Health-Behörden, besonders im Ausland und passenden akademischen Forschungsgruppen zu verbessern. Dies wird u.a. durch die Durchführung von Workshops, Hackathons und gemeinschaftlicher Open-Source-Software-Entwicklung realisiert.

5 Ethische/rechtliche Gesichtspunkte

Mit dem zu etablierenden *Ausbruchsinformations-System* werden vorhandene pseudonymisierte Daten, die gemäß IfSG an das RKI übermittelt werden, ausgewertet und die Ergebnisse zeitnah den Nutzenden am RKI, den Landesstellen und den Gesundheitsämtern präsentiert. Die Auswertung von Meldungsdaten erfolgt erst nach der Trennung der personenbezogenen Daten. Dadurch ergeben sich keine neuen ethischen bzw. rechtlichen Probleme.

Durch die zeitnahe Information der NutzerInnen im ÖGD ist z.B. in Ausbruchssituationen eine schnellere und zielgenauere Reaktion möglich.

Für die *Personalisierung* und die Bereitstellung der Informationen für Meldende werden Nutzereigneschaften bzw. Verhaltensdaten gespeichert, die entsprechende Sorgsamkeit erfordern. Hier sind entsprechende organisatorische und technische Maßnahmen zu ergreifen, die eine Einhaltung der Selbstbestimmungsrechte ermöglichen (Personalisierung als Opt-in). Auf die Beachtung des Datenschutzes wird zu jedem Zeitpunkt des Projektes strengstens geachtet.

Bei der Verwendung externer Datenquellen wird nur auf öffentlich zugängliche Ressourcen zurückgegriffen bzw. entsprechende Lizenzvereinbarungen getroffen.

6 Nutzen und Verwendung der Ergebnisse

6.1 Open-Data und Open-Access

Die Nachhaltigkeit des Projektes wird durch die Entwicklung eines frei zugänglichen *Ausbruchsinformations-Systems* erreicht. Dieses Informationssystem soll in den nächsten Jahren zentrales Werkzeug für die Ausbruchsbearbeitung im ÖGD werden. Teile der Funktionalität werden als *Dienste* (Web-Services) bereitgestellt, um so auch für weitere etablierte Anwendungen nutzbar gemacht zu werden.

Darüber hinaus sind die entwickelten Methoden für eine Vielzahl von Daten verwendbar. Neben den IfSG-Fällen können verschiedenste Datenquellen einbezogen werden, aus denen sich Zeitreihen extrahieren lassen. So z.B. die aktuellen Zahlen von Sterbefällen (Mortalitätsdaten), die Kommunikation über Krankheits-Symptome über dedizierte Kanäle oder in sozialen Medien (Syndromische Surveillance, M-Eco) oder weitere für die Epidemiologie wichtige Daten (Molekulare Surveillance, Antibiotika-Verbrauchszahlen, Wetterdienstdaten).

6.2 Einbindung der integrierten molekularen Surveillance (IMS)

Die *integrierte molekulare Surveillance* (IMS) ist Teil der Planungen von DEMIS und wird eine wesentliche Rolle in der zukünftigen epidemiologischen Surveillance spielen. Die großen Datenmengen und komplexen Zusammenhänge stellen sowohl ein großes Potential als auch eine große Herausforderung dar. Es werden adäquate Analyse-Werkzeuge und Visualisierungen eingesetzt, um die Erlangung wertvoller Erkenntnisse bezüglich Transmissionsketten und Ausbruchszugehörigkeit zu ermöglichen. Am Ende des Projekts sollen alle Teile des Systems (alle hier beschriebene Arbeitspakete) in der Lage sein, IMS Informationen zu berücksichtigen.

Eine enge Zusammenarbeit mit KollegInnen von FG36 (IMS der Tuberkulose) und NG4 (Bioinformatik) ist vorgesehen.

6.3 Einbindung der Antibiotika-Resistenz-Surveillance

Die Analyse der ARS[1]-Datenbank am RKI soll in naher Zukunft ebenso mit Hilfe von Algorithmen aus den Bereichen des *Maschinellen Lernens* und der *Netzwerk-Analyse* erfolgen. Die Ziele sind dabei unter anderem:

[1] Antibiotika-Resistenz-Surveillance

- die *zeitliche Entwicklung von AMR*[2] in Deutschland besser zu verstehen und
- Ursachen und Zusammenhänge aus komplexen, mehrschichtigen und heterogenen ARS-Groß-Daten zu identifizieren.

[2] Antimikrobiellen Resistenz

Dieses spezielle Thema und die dabei angewandten maßgeschneiderten Methoden sollen mittelfristig in die allgemeinen Dienste von DEMIS (insbes. des Informationsdienstes) eingebunden werden, um so einem breiten Nutzerkreis zur Verfügung zu stehen.

Eine enge Zusammenarbeit mit KollegInnen von FG37 (Surveillance von Antibiotikaresistenz und -verbrauch) und P4 (Epidemiologische Modellierung von Infektionskrankheiten) ist vorgesehen.

6.4 Einbindung der syndromischen Surveillance

Das RKI führt mehrere Projekte zur syndromische Surveillance durch, z.B. in Notaufnahmen (innerhalb von Krankenhäusern) oder in Unterkünften für Asylsuchende. Die im Rahmen des Projektes entwickelten Methoden sollen auch bei diesen Projekten zum Einsatz kommen.

Eine enge Zusammenarbeit mit KollegInnen von FG32 (Surveillance) und FG36 (Respiratorisch übertragbare Erkrankungen) im Rahmen der Projekte zur syndromischen Surveillance ist vorgesehen.

7 Arbeits- und Zeitplan

Die geplante Laufzeit des Projektes ist 3 Jahre: 01.01.2018 - 31.12.2020.

7.1 Arbeitspakete

Im Folgenden werden die einzelnen Arbeitspakete (AP) beschrieben. Für die Arbeit innerhalb der 9 Arbeitspakete sind entsprechende Teilaufgaben mit einer Schätzung in Personenmonaten (PM) angegeben. Wichtige Meilensteine (MS) und Deliverables (D) sind ebenso enthalten. Weiterhin werden die Abhängigkeiten zu den anderen Arbeitspaketen beschrieben und mögliche Risiken aufgeführt.

Eine Übersicht über die Verteilung der Arbeit auf die Personen bzw. auf die Projektlaufzeit (zusammen mit den Meilensteinen) ist in den Abschnitten Arbeitsverteilung bzw. GANTT zu finden.

Weiterentwicklung der Ausbruchserkennungsalgorithmen (AP-1, 18 PM)

- Entwicklung von aussagekräftigen Bewertungs-Methoden (Score, Ranking) (3 PM)
- Entwicklung von fachspezifischen Metriken für Variablen-Kategorien (6 PM)
- Optimierung der Verfahren zur Echtzeit-Tauglichkeit durch Vermeidung der bisher angewandten Brute-Force-Ansätze (9 PM, MS-1.1)
- Deliverables
  - Dokumentation der Methoden und Algorithmen incl. Vignette/User Guide (D-1.1)
- Abhängigkeiten
  - Keine, kann eigenständig bearbeitet werden.
- Risikoanalyse
  - Minimales Risiko der Nichterfüllung da alle notwendigen Vorarbeiten getätigt wurden und Wissen über mögliche realisierbare Ansätze vorhanden ist. Außerdem bestehen schon funktionierende Algorithmen, d.h. selbst bei geringem Erfolg können alle Folgeschritte weitergeführt werden.

Kofaktorenanalyse (AP-2, 18 PM)

- Auswahl, Analyse und Bereitstellung von Kontextinformationen aus dem Bereich strukturierter Daten (mind. 2) (2 x 6 PM, MS-2.1) und unstrukturierter Daten (6 PM, MS-2.2)
- Deliverables
  - Dokumentation des Workflows und der Metadaten (Struktur, Provenance (HCLS Community Profile)) (D-2.1 für strukturierte Daten, D-2.2 für unstrukturierte Daten)
- Abhängigkeiten

  - Methoden zur Extraktion von Informationen aus unstrukturierten Daten aus AP-3.
- Risikoanalyse
  - Geringes Risiko der Nichterfüllung da genügend interessante Datensätze bekannt sind. Selbst wenn keine geeigneten Methoden in AP-3 gefunden werden können weitere strukturierte Daten bereitgestelltwerden.

Datenaufbereitung via NLP und formaler Semantik (AP-3, 20 PM)

- Methodenauswahl zur Wissensextraktion aus nichtstrukturierten Daten (5 PM, MS-3.1)
- Einsatz einer NLP-Plattform (in Kooperation mit Uni Osnabrück) (15 PM)
- Deliverables
  - Konzept und Dokumentation (D-3.1)
- Abhängigkeiten
  - Keine, kann eigenständig bearbeitet werden
- Risikoanalyse
  - Mittleres Risiko der Nichterfüllung da vielversprechende Ansätze identifiziert wurden und mit dem Kooperationspartner (Institute of Cognitive Science, Universität Osnabrück) ein erfahrener Berater zur Seite steht. Trotzdem besteht die Gefahr, dass die extrahierten Information nicht für alle Quellen in vollem Umfang zu nützlichem Wissen verarbeitet werden kann. Es ist jedoch davon auszugehen, dass aus einigen der Quellen hilfreiche Informationen hervorgehen wird. Eine genauere Einschätzung wird durch eine Evaluation (AP-7) ermöglicht.

Werkzeuge zur Vorhersage und Simulation (AP-4, 16 PM)

- Vorhersage (MS-4.1)
  - Auswahl geeigneter Modelle zur Vorhersage von Zeitreihen ((S)ARIMA, ESM, GLM) (8 PM)
  - Parametrisierung der Modelle (4 PM)
  - Simulationen zur Bestimmung der Unsicherheit und Exploration von Szenarien (Markov-Chain-Monte-Carlo Ansatz) (4 PM)
- Deliverables
  - Dokumentation des Modells und Beschreibung des Verfahrens (Vignette/User Guide) (D-4.1)
- Abhängigkeiten
  - Ergebnisse zu Modellen und Methoden aus AP-1 sollen hier angewendet werden.
- Risikoanalyse
  - Geringes Risiko der Nichterfüllung, da verschiedene vielversprechende Ansätze zur Vorhersa-

ge und Simulation bereits ausfindig gemacht wurden. Selbst wenn keine großen Fortschritte in AP-1 gemacht werden, können Standardansätze verwendet werden und alle Folgeschritte können weitergeführt werden.

Personalisierung/Empfehlungssystem (AP-5, 14 PM)

- Inhaltsbasierte Empfehlungen (8 PM, MS-5.1)
- Kollaborative Filtermethoden (6 PM)
- Deliverables
  - Beschreibung des Verfahrens (Vignette/User Guide) (D-5.1)
- Abhängigkeiten
  - Keine, kann eigenständig bearbeitet werden.
- Risikoanalyse
  - Minimales Risiko der Nichterfüllung da die prinzipielle Machbarkeit im Rahmen einer Bachelorarbeit [9] in unserem Fachgebiet nachgewiesen wurde. Der genaue Grad des Nutzens der Empfehlung ist jedoch noch nicht abschätzbar. Durch die Aufteilung in inhaltsbasiert und kollaborative Methoden ist gewährleistet, dass sowohl von Anfang an sinnvolle Empfehlungen gegeben werden als auch mit jedem Feedback die Empfehlungen angepasst und verbessert werden können. Durch die Evaluation in AP-7 wird eine Bewertung ermöglicht.

Dashboard (AP-6, 22 PM)

- Bereitstellung des Ausbruchsinformations-Dashboards
  - Designprozess mit Nutzerbefragung (3 PM)
  - Visualisierung der Signale-Kennzahlen (5 PM, MS-6.1)
  - Integration der Kontextinformationen (7 PM, MS-6.1)
  - Integration der Methoden des Emfehlungssystems (7 PM, MS-6.2)
  - Integration der interaktiven Vorhersage zur Szenario-Erkundung (optional)
- Deliverables
  - User Guide des Dashboards incl. Screen-Cast (mind. 15 min) (D-6.1)
- Abhängigkeiten
  - Es sollen Ergebnisse aus allen vorangegangen APs (AP-1, . . . , AP-5) integriert werden.
- Risikoanalyse
  - Mittleres Risiko der Nichterfüllung da die Machbarkeit mit dem Prototyp eines funktionsfähigen Signal-Dashboards (Meldedaten + Signale) nachgewiesen wurde. Es besteht das Risiko,

dass einige der vorangegangen APs noch nicht ausreichend gute Ergebnisse liefern. In diesem Fall können die restlichen Ergebnisse bereits zusätzlich zu den Meldedaten und Signaldaten dargestellt werden.

Evaluation (AP-7, 13 PM)

- Evaluation der Analysemethoden zur Ausbruchsdetektion (Benchmark, 4 PM, MS-7.1)
- Evaluation der ausgewählten NLP-Plattform (5 PM, MS-7.2)
- Evaluation des Dashboards (incl. Kontextinformationen (UX, 4 PM, MS-7.3)
- Deliverables
  - Evaluationsberichte (incl. Evaluationsplan) (D-7.1, D-7.2, D-7.3)
- Abhängigkeiten
  - Es sollen Ergebnisse aus allen vorangegangenen APs evaluiert werden, wobei der Schwerpunkt auf den Arbeitspaketen AP-1, AP-3 und AP-6 liegt und letzteres die Evaluation von AP-2, AP-4 und AP-5 mit enthält.
- Risikoanalyse
  - Geringes Risiko der Nichterfüllung, da für die meisten Ergebnisse Evaluationsstrategien bereits identifziert wurden. Lediglich für die Ergebnisse aus AP-3 und AP-6 müssen neue Methoden der Evaluation getestet werden. Auch ohne zufriedenstellende Ergebnisse aus den vorhergenden APs kann ein Evaluationsbericht erstellt werden.

DEMIS-Integration (AP-8, 10 PM)

- Bezug und Aufbereitung von Kontext-Informationen (2 x 1 PM, 1 x 2 PM)
- Bereitstellung der Kontext-Information über eine öffentlich zugängliche Schnittstelle (als REST-API) (2 x 1 PM, MS-8.1)
- Bereitstellung der Methoden in R-Paketen (2 x 2 PM, MS-8.2, MS-8.3)
- Bereitstellung der NLP-Komponenten (optional)
- Deliverables
  - Beschreibung der Schnittstelle (z.B. in OpenAPI) (D-8.1)
  - Quellcode der zwei R-Pakete (bzw. Link zu CRAN/gitlab) (D-8.2, D-8.3)
- Abhängigkeiten
  - Es sollen Ergebnisse der Arbeitspakete AP-1 bis AP-5 in DEMIS integriert werden.
- Risikoanalyse
  - Geringes Risiko der Nichterfüllung, da bereits Erfahrung mit der Erstellung von R-Paketen und

der Beschreibung von Schnittstellen besteht. Außerdem wird das DEMIS Projekt im gleichen Fachgebiet entwickelt und die Zusammenarbeit mit den Entwicklern des DEMIS Projektes ist etabliert.

- Abgrenzung
  - Die Integration der Schnittstellen erfolgt im Rahmen des DEMIS-Projektes und ist nicht Gegenstand dieses Projektes. Ziel dieses Arbeitspaketes ist es, eine solche Integration gut vorzubereiten und entsprechende Schnittstellen und Dokumente zur Verfügung zu stellen.

Kommunikation/Dissemination (AP-9, 13 PM)

- Durchführung von Workshops/Hackathons zum fachlichen Austausch
  - Schwerpunkt: Kontextinformationen (3 PM, MS-9.1)
  - Schwerpunkt: Computerlinguistik (3 PM, MS-9.2)
  - Schwerpunkt: Ausbruchsinformation (3 PM, MS-9.3)
  - Dabei wird bei den Workshops/Hackathons für die Planung und Vorbereitung jeweils 1 PM angesetzt, für die Durchführung und Betreuung der Gäste jeweils 0.5 PM und für die Nach- und Aufbereitung jeweils 1.5 PM.
- Wiss. Veröffentlichungen (2 x 2 PM)
- Deliverables
  - Workshop-Bericht (als Teil des Zwischenberichts)
  - Einreichung von zwei Veröffentlichungen in peer-reviewed Journalen (D-9.1, D-9.2)
- Abhängigkeiten
  - Es sollen Ergebnisse der vorhergehenden APs präsentiert bzw. veröffentlicht werden.
- Risikoanalyse
  - Geringes Risiko der Nichterfüllung da bereits Erfahrungen in der Durchführung von Workshops und dem Anfertigen von wissenschaftlichen Veröffentlichungen bestehen. Zusätzlich kann der Kooperations Partner aus Osnabrück (Institute of Cognitive Science) extensive Erfahrung mit Hackathons auf dem Gebiet des Maschinellen Lernens vorweisen und hat uns ihre Zustimmung bei der Durchführung solcher Hackathons bereits zugesagt. Die Erfolgsaussichten der wissenschafltichen Veröfftlichungen hängen natürlich vom Erfolg der vorhergehenden APs ab. Es ist jedoch davon auszugehen, dass zumindest einige der APs interessante Erkentnisse liefern werden. Der Erfolg von Hackathons und Workshops ist ebenfalls schlecht abschätzbar. Bisher wurden sehr positive Erfahrungen mit beiden Verfahren gemacht (siehe auch Workshop-Bericht als Anlage).

Arbeitsverteilung

Die Tablle 1 gibt eine Übersicht zur Verteilung der Arbeitspakete auf die Projektmitarbeitenden. Als Einheit wurden Personenmonate gewählt. Die Abkürzungen sind in Abschnitt 9.1 erklärt. Neben der Verteiung für die WissenschaftlerInnen und die ProgrammiererIn wurden auch die studentischen Hilfskräfte berücksichtigt.

AP	W1	W2	W3	P1	Summe	S1	S2
AP-1 Methoden	2	8	4	4	18		
AP-2 Kontextinformationen	8	2	4	4	18	12	
AP-3 NLP	4	4	8	4	20	12	12
AP-4 Vorhersage	3	3	4	6	16		9
AP-5 Personalisierung	6		4	4	14	6	
AP-6 Dashboard	3	4	6	9	22		6
AP-7 Evaluation	3	8	1	1	13	3	3
AP-8 Integration	2	2	3	3	10		
AP-9 Dissemination	5	5	2	1	13	3	6
Summe	36	36	36	36	144	36	36

Tabelle 1: Übersicht zur Arbeitsverteilung (PM) auf die Projektmitarbeitenden

GANTT

In Abbildung 1 ist ein Übersicht zur zeitlichen Verteilung der Arbeitspakete und der Meilensteine gegeben. Wir haben hier auf die Nummerierung der Meilensteine verzichtet und stattdessen sprechende Namen gewählt. Eine inhaltliche Zuordnung wird dadurch erleichtert.

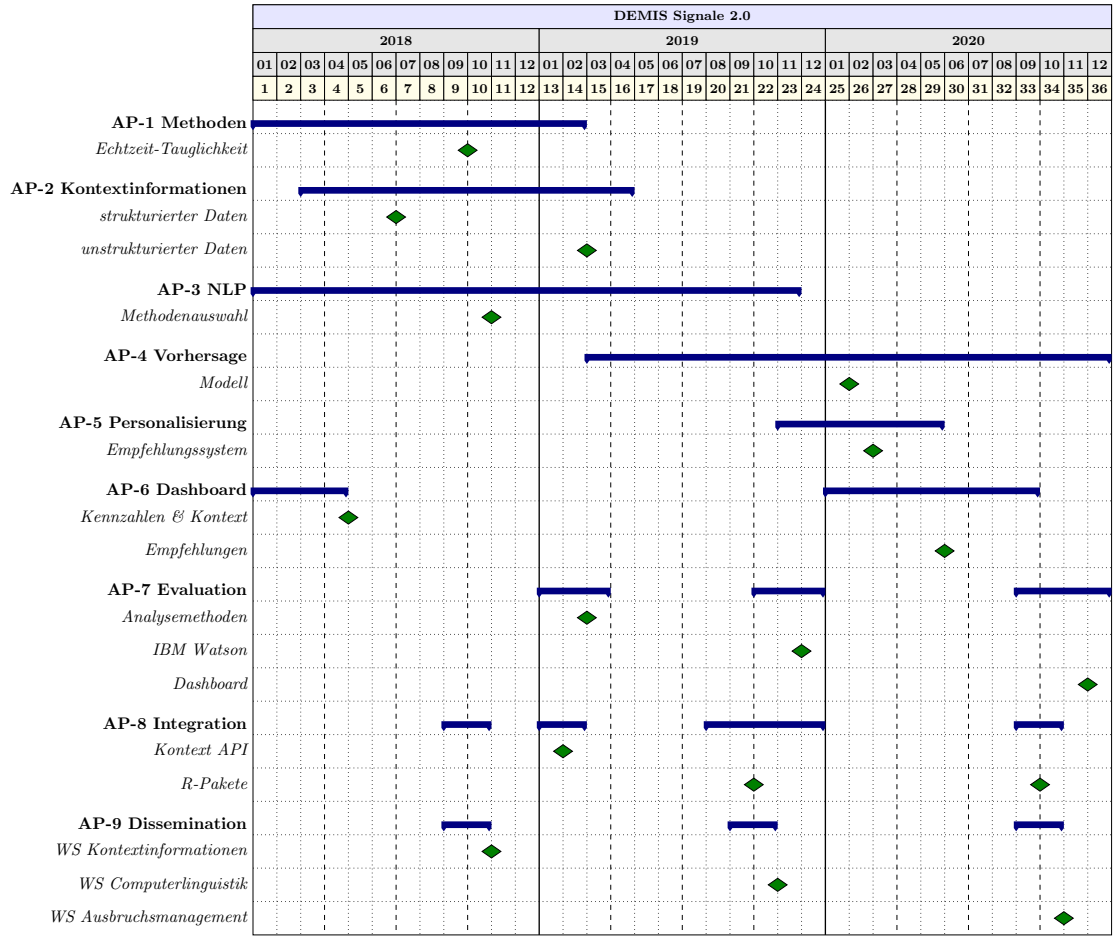

Abbildung 1: Übersicht über die Arbeitspakete und ihre Meilensteine

7.2 Zeitplan

Erstes Jahr (2018)

- Entwicklung und Evaluation der Methoden (W1, W2)[3]
- Implementierung von Bewertungs-Methoden (W3, P1)
- Bereitstellung des Ausbruchsinformations-Dashboards (W1, W3, P1)
- Bereitstellung von Kontextinformationen (W1, W3, P1)
- Architektur und Prototyping der NLP-Komponenten (W2, W3, P1, S2)
- Organisation eines Workshops/Hackathons zum Thema Signalerkennung mit und Vertretern anderer Public-Health-Institute und Universitäten (Schwerpunkt: Kontextinformationen) (W1, W3, S1)(MS-9.1)

[3] für Abkürzungen siehe 9.1

Zweites Jahr (2019)

- Umsetzung der Verfahren als Web-API und open-source Bibliotheken (W3, P1)
- Auswahl und Entwicklung von Vorhersagemethoden (W1, W2, P1)
- Umsetzung und Evaluation der NLP-Komponenten (W2, W3, P1, S2)
- Publikation der Ergebnisse (W1, W2, W3)
- Organisation eines Workshops/Hackathons zum Thema Ausbruchserkennung mit Wissenschaftlern und Vertretern anderer Public-Health-Institute und Universitäten (Schwerpunkt: Computerlinguistik) (W2, W3, S2)(MS-9.2)

Drittes Jahr (2020)

- Integration der Vorhersage in das Ausbruchsinformations-Dashboard (W2, W3, P1)
- Umsetzung des Recommender-Systems zur Personalisierung (W1, W3, P1, S1)
- Bereitstellung des Ausbruchsinformations-Dashboards mit Personalisierung und NLP-Komponenten (W3, P1, S2)
- Organisation eines Workshops/Hackathons zum Thema Ausbruchserkennung mit Wissenschaftlern und Vertretern anderer Public-Health-Institute und Universitäten (Schwerpunkt: Ausbruchsinformations-System) (W1, W2, S1)(MS-9.3)

7.3 Kooperationspartner

- Universität Osnabrück, TU Darmstadt, HPI Potsdam
  - Methodenentwicklung in den Bereichen *Machinelles Lernen*, *Computerlinguistik (NLP)* und *Künstliche Intelligenz*
- National Infection Service Public Health England
  - Methodenentwicklung in Statistik, fachliche Expertise
- Landesstellen, Gesundheitsämter
  - Fachliche Expertise und Pilotnutzung

Als Anhang sind entsprechende *Letters of Intent* beigefügt. In Tabelle 2 sind ausgewählte Ansprechpartner aufgeführt.

Tabelle 2: Ausgewählte Ansprechpartner

Name	**Organisation**	**Ort**	**Fachgebiet**
Gomes Dias , Joana	ECDC	Solna, SE	Statistik
Khnafo, Dalhia	epiconcept	Paris, FR	Software
Orchard, Francisco	epiconcept	Paris, FR	Data science
Vallée, Morgane	epiconcept	Paris, FR	Statistik
Raimbault, Bruno	Freelance	Barcelona, ES	Visualisierung
Lytras, Theodore	HCDCP	Athens, GR	Statistik
Jombart, Thibaut	Imperial College	London, UK	Statistik
Nouvellet, Pierre	Imperial College	London, UK	Statistik
Höhle, Michael	IQTIG	Berlin, DE	Statistik
Schipper, Maarten	RIVM	Utrecht, NL	Statistik
van de Kassteele, Jan	RIVM	Utrecht, NL	Statistik
Rohde, Martin	OFFIS	Oldenburg, DE	Mathematik
Kühnberger, Kai-Uwe	Osnabrück University	Osnabrück, DE	ML, AI
Nieters, Pascal	Osnabrück University	Osnabrück, DE	ML, AI
Elliot, Alex	Public Health England	London, UK	Epidemiologie
Morbey, Roger	Public Health England	London, UK	Statistik
Le Strat, Yann	Santé publique France	Saint-Maurice, FR	Data science
Bjelkmar, Pär	The PH Agency of Sweden	Solna, SE	Statistik
Källberg, Henrik	The PH Agency of Sweden	Solna, SE	Statistik
Loza Mencía, Eneldo	TU Darmstadt	Darmstadt, DE	ML, AI
Colón-González, Felipe	University of East Anglia	Norwich, UK	Statistik
Noufaily, Angela	University of Warwick	London, UK	Statistik
Corberán-Vallet, Ana	University Valencia	Valencia, ES	Statistik

8 Risikofaktoren

Das größte Risiko besteht in der zeitigen Gewinnung qualifizierter MitarbeiterInnen. Da bei diesem Projekt eine interdisziplinäre Arbeitsweise benötigt wird, ist mit einer entsprechenden Einarbeitungszeit zu rechnen. Gelingt es, erfahrene oder bereits bewährte WissenschaftlerInnen einzustellen, kann dieses Risiko weitgehend minimiert werden.

Auf Grund des Umfangs des Projektes besteht ein mittleres Risiko der Verzögerung einzelner Systemteile. Durch umfangreiche Vorarbeiten wurde gezeigt, dass das prinzipielle Vorgehen zu nutzbaren und von den NutzerInnen erwarteten Ergebnissen führt. Im Verlauf des Projektes ist geplant, dass das schon im Einsatz befindliche System permanent ergänzt wird und durch eine stetige Rückkopplung mit den Nutzenden kontinuierlich verbessert wird (Agile Vorgehensweise). Zudem sind die Arbeitspakete soweit entkoppelt, dass keine größeren Verzögerungen auf Grund von Verzögerungen oder Misserfolg in anderen APs entstehen sollten.

Da moderne Methoden auf neue Bereiche angewendet werden ist vorher nicht möglich genau abzuschätzen wie hoch der Nutzen der einzelnen Ergebnisse ist. Durch die Vorarbeiten, Erfahrungen und Kompetenzen der Mitarbeiter des Fachgebiets und der unterstützenden Kooperationen in den betreffenden Gebieten sowohl der Forschung, Entwicklung als auch Umsetzung ist das Risiko gering, dass das Projekt zu keinen nützlichen Produkten führt. Selbst bei einem geringfügigem Erfolg in einzelnen APs kann trotzdem noch ein wertvolles Produkt entstehen.

Risiken von Kooperationen werden dadurch minimiert, dass auf vertragliche Grundlagen geachtet wird, die eine Nachnutzung der Ergebnisse durch das RKI ermöglichen. Ebenso wird auf die Verwendung weitverbreiteter und zukunftstauglicher Technologien geachtet.

Eine ausführliche Risikoanalyse ist in der Beschreibung der einzelnen Arbeitspakete zu finden.

9 Finanzplan

Die geplante Laufzeit des Projektes ist 3 Jahre: 01.01.2018 - 31.12.2020.

9.1 Personal

- 3 Wiss. Mitarbeiter (TVöD 13/14)
  - 2 WM mit statistischer Expertise (W1, W2)
  - 1 WM mit fundierten IT-Kenntnissen (W3)
- 1 Programmierer (TVöD 10/11) (P1)
- 2 Studentische Hilfskräfte (40 Stunden/Monat) (S1, S2)

9.2 Sachmittel

- Mittel für 3 Workshops/Hackathons für jeweils 20 Personen à 1.000€
  - Raummiete
  - Drucksachen
  - Reisekosten
  - Moderation
- Sachmittel für Lizenzen oder Daten ca. 10.000€ pro Jahr
  - externe Aufträge
  - Sonstige Sachausgaben
- Generelle Sachmittel (Reisekosten, Bewirtung etc.) ca. 10.000€ pro Jahr

Sachausgaben	**Betrag**
Raummiete	2000
Geräte und Ausstattungsgegenstände	4000
Drucksachen und Büromaterial	1000
Aufwendungen für (Dienst-) Reisen	20000
Vergabe von Aufträgen	8000
Post- und Fernmeldegebühren	0
Sonstige Sachausgaben	5000
Summe	40000

Tabelle 3: Sachmittel pro Jahr

9.3 Gesamtübersicht

	Gruppierung	Faktor	Satz	Summe
Kalenderjahr 2018				399878
Wissenschaftlicher Mitarbeiter	E 14	3	90313	270939
Programmierung	E 11	1	78379	78379
Studentische Hilfskräfte		2	5280	10560
Sachmittel		1	20000	20000
Organisation eines Workshops		1	20000	20000
Kalenderjahr 2019				406863
Wissenschaftlicher Mitarbeiter	E 14	3	92119	276357
Programmierung	E 11	1	79946	79946
Studentische Hilfskräfte		2	5280	10560
Sachmittel		1	20000	20000
Organisation eines Workshops		1	20000	20000
Kalenderjahr 2020				413988
Wissenschaftlicher Mitarbeiter	E 14	3	93961	281883
Programmierung	E 11	1	81545	81545
Studentische Hilfskräfte		2	5280	10560
Sachmittel		1	20000	20000
Organisation eines Workshops		1	20000	20000
Gesamt				1.220.729

Tabelle 4: Mittelübersicht

Literatur

[1] Steven Bird, Ewan Klein und Edward Loper. *Natural Language Processing with Python*. Bd. 43. 2009, S. 479. ISBN: 9780596516499. DOI: 10.1097/00004770-200204000-00018. arXiv: arXiv:1011.1669v3. URL: http://www.amazon.com/dp/0596516495.

[2] Dirk Brockmann und Dirk Helbing. „The Hidden Geometry of Complex, Network-Driven Contagion Phenomena“. In: *Science* 342.6164 (Dez. 2013), S. 1337–1342. ISSN: 0036-8075. DOI: 10.1126/science.1245200. URL: http://www.sciencemag.org/cgi/doi/10.1126/science.1245200.

[3] DelphiCast Team. *ILI Nearby*. URL: http://delphi.midas.cs.cmu.edu/nowcast/about.html.

[4] Kerstin Denecke u. a. „Event-Driven Architecture for Health Event Detection from Multiple Sources“. Englisch. In: *Proceedings of the XXIII International Conference of the European Federation for Medical Informatics (MIE 2011)*. Oslo, NO: IOS Press, 2011, S. 160–164. ISBN: 978-1-60750-805-2. URL: http://www.fit.vutbr.cz/research/view_pub.php?id=9614.

[5] Mark Dredze u. a. „HealthTweets.org: A Platform for Public Health Surveillance using Twitter“. In: (). URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.431.1469.

[6] D Faensen u. a. „SurvNet@RKI–a multistate electronic reporting system for communicable diseases.“ In: *Euro surveillance* 11.4 (2006), S. 100–3. ISSN: 1560-7917. URL: http://www.ncbi.nlm.nih.gov/pubmed/16645245.

[7] David C. Farrow u. a. „A human judgment approach to epidemiological forecasting“. In: *PLOS Computational Biology* 13.3 (März 2017). Hrsg. von Samuel Alizon, e1005248. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005248. URL: http://dx.plos.org/10.1371/journal.pcbi.1005248.

[8] Stéphane Ghozzi, Alexander Ullrich und Göran Kirchner. „Early Recognition of Infectious Disease Outbreaks“. In: *SIREMTI*. 2017.

[9] Danilo Günzel. „Konzeption und Implementierung eines Empfehlungssystems zur personalisierten Aufbereitung von infektionsepidemiologischen Daten“. Bachelor-Thesis. Hochschule für Wirtschaft und Recht Berlin, 2017.

[10] IBM. *IBM - Watson - Deutschland*. URL: http://www-05.ibm.com/de/watson/.

[11] Dietmar Jannach u. a. *Recommender Systems: An Introduction*. Cambridge, UK: Cambridge University Press, 2010. ISBN: 978-0-521-49336-9.

[12] Gerard Krause u. a. „SurvNet Electronic Surveillance System for Infectious Disease Outbreaks, Germany.“ In: (2007). Artikel; published; Centers for Disease Control and Prevention; http://www.cdc.gov/eid/content/13/10/1548.htm; Emerging Infectious Diseases; 13; 2007; 10. URL: http://edoc.rki.de/docviews/abstract.php?id=347.

[13] Martin Kulldorff u. a. „A Space–Time Permutation Scan Statistic for Disease Outbreak Detection“. In: *PLoS Medicine* 2.3 (Feb. 2005). Hrsg. von Sally M. Blower, e59. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.0020059. URL: http://dx.plos.org/10.1371/journal.pmed.0020059.

[14] Mykhailo Lobur, Andriy Romanyuk und Mariana Romanyshyn. „Using NLTK for educational and scientific purposes“. In: *11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM 2011)* (2011), S. 426–428.

[15] Sebastian Meyer, Leonhard Held und Michael Höhle. „Spatio-Temporal Analysis of Epidemic Phenomena Using the R Package surveillance“. In: (Nov. 2014). arXiv: 1411.0416. URL: http://arxiv.org/abs/1411.0416.

[16] Ramesh Nallapati u. a. „Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond“. In: (Feb. 2016). arXiv: 1602.06023. URL: http://arxiv.org/abs/1602.06023.

[17] Xin Pan und Peter Liu. *Sequence-to-Sequence with Attention Model for Text Summarization*. URL: https://github.com/tensorflow/models/tree/master/textsum.

[18] Alexander M. Rush, Sumit Chopra und Jason Weston. „A Neural Attention Model for Abstractive Sentence Summarization“. In: (Sep. 2015). arXiv: 1509.00685. URL: http://arxiv.org/abs/1509.00685.

[19] Marcel Salathé u. a. „Digital Epidemiology“. In: *PLOS Computational Biology* 8.7 (Juli 2012), S. 1–3. DOI: 10.1371/journal.pcbi.1002616. URL: https://doi.org/10.1371/journal.pcbi.1002616.

[20] Maëlle Salmon, Dirk Schumacher und Michael Höhle. „Monitoring count time series in R: Aberration detection in public health surveillance“. In: *Journal of Statistical Software* 70.1 (2016), S. 1–35. ISSN: 1548-7660. DOI: 10.18637/jss.v070.i10. arXiv: 1411.1292. URL: https://www.jstatsoft.org/article/view/v070i10.

[21] Maëlle Salmon u. a. „A system for automated outbreak detection of communicable diseases in Germany“. In: *Eurosurveillance* 21.13 (2016). ISSN: 15607917. DOI: 10.2807/1560-7917.ES.2016.21.13.30180.

[22] Maëlle Salmon u. a. „Bayesian outbreak detection in the presence of reporting delays“. In: *Biometrical Journal* 57.6 (2015), S. 1051–1067. ISSN: 15214036. DOI: 10.1002/bimj.201400159.

[23] Adam Spannbauer. *lexRankr: Extractive Summarization of Text with the LexRank Algorithm*. URL: https://cran.r-project.org/web/packages/lexRankr/index.html.

[24] Sean J Taylor und Benjamin Letham. „Forecasting at Scale“. In: (2017). URL: https://facebookincubator.github.io/prophet/static/prophet_paper_20170113.pdf.

[25] Xue-Wen Chen und Xiaotong Lin. „Big Data Deep Learning: Challenges and Perspectives“. In: *IEEE Access* 2 (2014), S. 514–525. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2014.2325029. URL: http://ieeexplore.ieee.org/document/6817512/.

[26] Hui Zou und Trevor Hastie. „Regularization and variable selection via the elastic net“. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 67.2 (2005), S. 301–320. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2005.00503.x.