Kurzbericht zum BMG-geförderten Forschungsvorhabens

Vorhabentitel	DEMIS Signale 2.0
Schlüsselbegriffe	Surveillance, Infektionskrankheit, Ausbruchserkennung, Machine Learning
Vorhabendurchführung	Abteilung für Infektionsepidemiologie, Robert Koch-Institut
Vorhabenleitung	Dr. Hermann Claus, Dr. Alexander Ullrich
Autor(en)/Autorin(nen)	Dr. Hermann Claus, Dr. Alexander Ullrich
Vorhabenbeginn	01.12.2017
Vorhabenende	31.12.2021

1. Vorhabenbeschreibung und Vorhabenziele

Die Überwachung von Infektionskrankheiten ist eine der Kernaufgaben des Robert Koch-Instituts (RKI). Zu diesem Zweck werden über verschiedene Surveillancesysteme Daten über das Auftreten von Infektionskrankheiten in der deutschen Bevölkerung erhoben. Das Infektionsschutzgesetz (IfSG) beschreibt die Meldepflicht für viele Infektionskrankheiten bzw. Erreger. Die von den Gesundheitsämtern verarbeiteten Daten werden elektronisch über die Software SurvNet [1] an das RKI übermittelt.

Die über das Meldesystem erhobenen Daten werden von Experten und Expertinnen am RKI und in den zuständigen Landesbehörden auf außergewöhnliche Trends und mögliche Ausbruchsereignisse hin untersucht. Die Vielzahl von zu überwachenden Infektionskrankheiten, Regionen, Risikogruppen und demografischen Gruppen schafft die Notwendigkeit von automatisierten Prozessen zur Unterstützung für die Erkennung besonderer Ereignisse. Im Vorgängerprojekt „DEMIS Signale“ wurde dazu ein automatisiertes Frühwarnsystem entwickelt [2], das mit Hilfe statistischer Methoden [3] Auffälligkeiten in den Meldedaten, die sogenannten Signale, findet und in täglich erstellten Berichten aufbereitet und zur Verfügung stellt.

Ziel dieses Projekts ist es, neuartige Methoden des maschinellen Lernens zu erforschen, die zum einen andere Datenquellen extrahieren und integrieren als auch weitere räumliche und zeitliche Effekte in der Dynamik von Infektionskrankheiten berücksichtigen. Die Ergebnisse der entwickelten Methoden sowie die Meldedaten und externe Daten sollen in einem Dashboard so

dargestellt werden, dass Epidemiologinnen und Epidemiologen einen umfassenden Eindruck der aktuellen epidemiologischen Lage erhalten und gezielt erforschen können.

2. Durchführung und Methodik

Auf den Gebieten des maschinellen Lernens und der Infektionsepidemiologie werden vielversprechende Datenquellen und neuartige Methoden identifiziert, die es erlauben, zusätzliche Erkenntnisse über Verlauf und Dynamik von Infektionskrankheiten zu gewinnen.
Um entscheiden zu können, welche dieser Methoden und Datenquellen geeignet sind, um die etablierten Methoden, wie z. B. dem Erkennen von Ausreißern in Zeitreihen mit Hilfe einfacher Regressionsverfahren, zu ergänzen oder abzulösen, müssen die Methoden evaluiert werden. Dafür werden aus realen Daten sowie Simulationen sogenannte Benchmarking-Datensätze erstellt. Dies sind Datensätze, bei denen das korrekte Auftreten und Ausdehnung der Ausbrüche bekannt ist und womit die erfolgreiche Erkennung von Ausbrüchen, aber auch das Verpassen von Ausbrüchen oder falschen Alarmen bestimmt werden kann. Mittels geeigneter Metriken, z. B. einer Falsch-Positiven-Rate, können mit den erstellten Datensätze Ausbruchserkennungsmethoden innerhalb eines Evaluierungs-Frameworks systematisch bewertet werden. Dies ist ein System, welches sicherstellt, dass die Ergebnisse der Metriken für alle verwendeten Methoden auch vergleichbar sind.
Sowohl beim Entwurf als auch bei der Evaluation der verschiedenen Teilschritte werden Nutzerinnen und Nutzer, also Epidemiologinnen und Epidemiologen des Öffentlichen Gesundheitsdienst (ÖGD) und andere Experten und Expertinnen, eingebunden. Dafür werden Fachkreise durchgeführt, Netzwerktreffen arrangiert und Publikationen angefertigt.

3. Gender Mainstreaming

Bei der Analyse von Meldedaten werden neben anderen Parametern (z.B. Alter, Subtyp des nachgewiesenen Erregers) auch Angaben zum Geschlecht (männlich, weiblich, divers, unbekannt) berücksichtigt. In den Dashboards zu den Meldedaten werden geeignete Darstellungen verwendet, die die Häufigkeiten nach allen Angaben zu Geschlecht stratifiziert darstellen. Allerdings erfassen nicht alle Surveillancesysteme das Geschlecht bzw. kann nicht bei allen Datenquellen eine getrennte Betrachtung nach Geschlecht durchgeführt werden.

4. Ergebnisse, Schlussfolgerung, Fortführung

Es wurden mehrere neue Methoden zur Analyse von Infektionskrankheiten entwickelt, evaluiert und veröffentlicht. Zur Vorhersage von Fallzahlen unter Berücksichtigung von raumzeitlicher Dynamik und weiteren soziodemografischen und geografischen Faktoren haben wir eine bayesianische Methode entwickelt, welche eine vergleichbare Performance wie andere state-of-the-art Vorhersagemethoden [4] zeigt, aber eine bessere Aussage über die Unsicherheit der Vorhersage

erlaubt. Dies wird durch das Hinzufügen von Expertenwissen über bekannte Quellen der Unsicherheit, z. B. der zeitlichen Verteilung von Ausbruchsfällen, erzielt. Zur Ausbruchserkennung auf Einzelfallebene haben wir eine semi-überwachte Clusteringmethode (Identifizieren von Gruppen ähnlicher Fälle) mit einem modernen Dimensionsreduktionsansatz (Transformation hochdimensionaler Daten in zwei Dimensionen, welche dann auch visuell dargestellt werden können) verbunden, welches viel weitergehendere Interpretationen als konventionelle Zeitreihen-basierte Methoden erlaubt. Für die systematische Evaluierung von Ausbruchserkennungsalgorithmen haben wir ein Framework entwickelt welches die Vergleichbarkeit sicher stellt und nach mehreren Metriken, wie z. B. Präzision, F1-Score, Falsch-Positiven-Rate (FPR) und einer gewichteten Version des FPR, beurteilt. Dies ermöglicht außerdem eine Optimierung der Parameterwahl.
Des Weiteren haben wir Methoden entwickelt, um neue Informationsquellen zu erschließen. Zur Nutzung von Textmitteilungen über Quellen wie ProMED und WHO Outbreak News haben wir eine Natural Language Processing (NLP)-Pipeline (computergestützter Prozess zur Verarbeitung natürlicher Sprache) entwickelt, welche wichtige Informationen extrahiert, die Relevanz der Mitteilung einschätzt und die Ergebnisse in einer relationalen Datenbank und über eine Programmierschnittstelle sowie für ein interaktives Dashboard zur Verfügung stellt. Dies findet Anwendung im Screening der Ereignis-basierten Überwachung von internationalen Nachrichten.
Zur Erschließung von Social-Media-Datenquellen haben wir ein Programm entwickelt, welches automatisiert die Schnittstellen von Wikipedia und Twitter abfragt, nach relevanten Schlagworten für verschiedene Infektionskrankheiten durchsucht und die Suchergebnisse aggregiert und persistiert. Daraus können Zeitreihen erstellt werden, die mit tatsächlichen Fallzahlen der jeweiligen Infektionskrankheit verglichen werden. Diese Methode ist für einige der Infektionskrankheiten, z. B. die respiratorischen Erreger Influenza und RSV sowie die gastro-intestinalen Erreger *Campylobacter* und Salmonellen. realisiert, für die bereits teils umfangreichere Surveillancedaten zur Verfügung stehen.
Die Ergebnisse der verschiedenen Methoden und Analysen werden zusammen mit Daten aus verschiedenen Surveillancesystemen in eigens entwickelten Dashboards zur Verfügung gestellt. Diese werden RKI-intern z.B. zur Einschätzung der Schwere von Grippewellen, der Untersuchung von Zusammenhängen zwischen Wetter und vektor-basierten Krankheiten und der Ausbruchserkennung von allgemeinen Infektionskrankheiten verwendet. Während der Pandemie wurden weitere automatisierte Berichte und Dashboards entwickelt, die sowohl von Fachgremien als auch der Öffentlichkeit genutzt werden. Dabei floss die Erfahrung aus der nutzerorientierten Entwicklung der internen Anwendungen mit ein.
Im Rahmen des Projekts fand auch eine starke internationale Vernetzung auf dem Gebiet der Ausbruchserkennung statt, die in verschiedenen Formaten über die Laufzeit des Projekts fortgeführt wird.
Die Früherkennung von Ausbrüchen von Infektionskrankheiten am RKI ist durch die entwickelten Tools, Berichte und Dashboards nachhaltig gestärkt und wird im Rahmen weiterer Projekte sowie

der ÖGD-Kontaktstelle fortgesetzt. Außerdem wurde ein erheblicher Beitrag zur Forschung auf dem Gebiet der Datenanalyse von Infektionskrankheiten in Form von öffentlichem Programmiercode und Veröffentlichungen in Fachzeitschriften geleistet.

Masterarbeiten:

- Kohn, K.J. **Individual-based detection of infectious disease outbreaks using machine learning techniques.** 2020. Christian-Albrechts-Universität Kiel
- Wagner, B. **Classifying Emergency Department Data to Improve Syndromic Surveillance: From Mixed Data Types to ICD Codes and Syndromes.** 2020. Universität Bielefeld
- Lison, A. **Interpretable Hierarchical Forecasting of Infectious Diseases.** 2020, WWU Münster, Universität Münster
- Becker, F. **Use of Social-Media Data in an Epidemiological Context.** 2020, Beuth Hochschule für Technik Berlin
- Abbood, A. **Automatic Information Extraction and Relevance Evaluation of Epidemiological Texts Using Natural Language Processing**. 2019. Universität Osnabrück
- Busche, R. **Systematic Evaluation and Optimization of Outbreak-Detection Algorithms Based on Labeled Epidemiological Surveillance Data**. 2019. Universität Osnabrück

Veröffentlichungen in Fachzeitschriften:

- Stojanović O, Leugering J, Pipa G, Ghozzi S, Ullrich A. **A Bayesian Monte Carlo approach for predicting the spread of infectious diseases.** PLoS ONE 14 (12): e0225838. Epub Dec 18. https://doi.org/10.1371/journal.pone.0225838
- Abbood A, Ullrich A, Busche R, Ghozzi S (2020) **EventEpi—A natural language processing framework for event-based surveillance.** PLOS Computational Biology 16(11): e1008277. https://doi.org/10.1371/journal.pcbi.1008277
- Sarma N, Ullrich A, Wilking H, Ghozzi S, Lindner AK, Weber C, Holzer A, Jansen A, Stark K, Vygen-Bonnet S. **Surveillance on speed: Being aware of infectious diseases in migrants mass accommodations – an easy and flexible toolkit for field application of syndromic surveillance, Germany, 2016 to 2017.** Euro Surveill. 23 (40): pii=1700430. https://doi.org/10.2807/1560-7917.ES.2018.23.40.1700430
- Ullrich, Alexander et al. **Impact of the COVID-19 pandemic and associated non-pharmaceutical interventions on other notifiable infectious diseases in Germany: An analysis of national surveillance data during week 1–2016 – week 32–2020.** The Lancet Regional Health – Europe, Volume 6, 100103. https://doi.org/10.1016/j.lanepe.2021.100103
- Bracher, J., Wolffram, D., Deuschel, J., Ullrich A. et al. **A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave**. Nat Commun 12, 5173 (2021). https://doi.org/10.1038/s41467-021-25207-0

5. Umsetzung der Ergebnisse durch das BMG

Das BMG hat auf Bundesebene die Zuständigkeit für die medizinischen und rechtlichen Fragen des Infektionsschutzes. Die Entwicklung von neuen praxisrelevanten Methoden und der Einsatz von innovativen Informationssystemen im RKI bilden wichtige Elemente einer Gesamtstrategie zur Abwehr und schnellen Bewältigung von Infektionsgefahren. Das vorliegende Vorhaben wird dazu einen wichtigen Beitrag leisten. Im Ausblick werden die Erkenntnisse und Produkte in das geplante Informationssystem im Rahmen der Weiterentwicklung des elektronischen Melde- und Informationssystems (DEMIS) einfließen. Während der Pandemie wurde der Nutzen von erweiterten Methoden der epidemiologischen Modellierung zur Vorhersage und Szenarienmodellierung deutlich, genauso wie die Integration verschiedener Surveillancesysteme und Kooperation mit externen Forschungsgruppen. Hier besteht Bedarf an der Vernetzung und weiterer Forschung.

Tabelle 1*: Übersicht der internen und öffentliche Dashboards und Berichte die während des Signale Projekts entstanden sind.*

	Beschreibung	Zielgruppe
A	Deutsch-Polnischer Forecast Hub: Plattform auf der Forschungsgruppen Vorhersage Ergebnisse einreichen und darstellen lassen können	Öffentlichkeit
B	ECDC Forecast Hub: Analog zu A, für ECDC Mitgliedsländer	Öffentlichkeit
C	Täglicher Lagebericht des RKI: Zusammenfassung relevanter Surveillancedaten zur COVID-19-Pandemie	Öffentlichkeit
D	Pandemieradar: Übersicht der relevantesten Indikatoren der COVID-19-Pandemie	Öffentlichkeit
E	Signale-Berichte: Präsentation der Ausbruchserkennungsergebnisse für 32 Krankheiten	Intern und interessierte Landesbehörden
F	Signale-Dashboard: interaktive Darstellung angelehnt an die Signale Berichte	Intern
G	COVID-19-Signale-Dashboard: Interaktive Darstellung von Ausbruchserkennung auf COVID-19-Daten verschiedener Surveillancesysteme	Intern
H	FSME-Dashboard: Tool zur Erkundung externer Einflüsse (z. B. Wetter) auf durch Zecken übertragene Erreger wie FSME und Borreliose	Intern
I	Vacmap: Gemeinsame Präsentation von Meldedaten und Impfquoten für Masern und Rotaviren	Intern
J	Influenza-Dashboard: Darstellung der verschiedenen Datenquellen für Influenza zur Einschätzung nach den PISA-Guidelines der WHO	Intern

Abbildung 1: Öffentliche Dashboards und Berichte

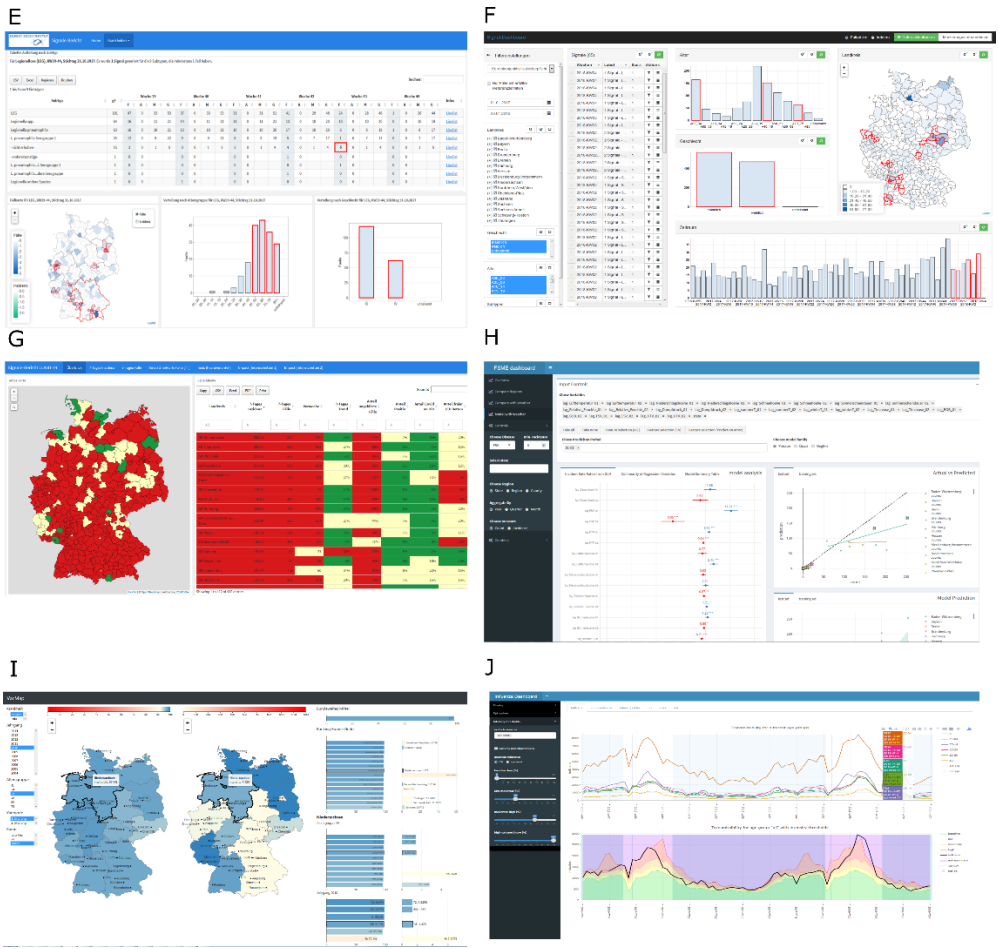

Abbildung 2: Interne Dashboards und Berichte

Verwendete Literatur

1. KRAUSE, G.; ALTMANN, D.; FAENSEN, D.; et al. **SurvNet electronic surveillance system for infectious disease outbreaks, Germany**. *Emerg Infect Dis*. 2007; 13(10):1548-1555
2. SALMON, M.; SCHUMACHER, D.; BURMANN, H.; FRANK, C.; CLAUS, H.; HÖHLE, M. **A system for automated outbreak detection of communicable diseases in Germany**. Euro Surveill. 2016; 21: 13
3. NOUFAILY, A.; ENKI, D.G.; FARRINGTON, P.; GARTHWAITE, P.; ANDREWS, N.; CHARLETT, A. **An improved algorithm for outbreak detection in multiple surveillance systems**. Statist. Med. 2013; 32: 1206-1222.
4. MEYER, S.; HELD, L.; HÖHLE, M. **Spatio-temporal analysis of epidemic phenomena using the R package surveillance**. Journal of Statistical Software 2017; 77(11): 1–55