Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Analyse von Visitentexten mittels maschinellem Lernen zur Optimierung der Leitlinienadhärenz auf einer Intensivstation

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von: Martin Hoffmann
geboren am: 06.09.1998
geboren in: Berlin

Gutachter: Prof. Dr. med. Dr. rer. nat. Felix Balzer
Dr. med. Fridtjof Schiefenhövel

eingereicht am: ..... verteidigt am: .....

Zusammenfassung

Hier kommt ein toller Abstract

Inhaltsverzeichnis

1 Einführung

1.1 Datenerfassung auf Intensivstationen

blabla. 50% der Fehler im Krankenhaus werden auf solche Kommunikationsprobleme zurückgeführt (Bhasale et al. 1998) und gelten als wichtiger Faktor für erhöhte Krankenhausmortalitätsraten (Wilson et al. 1995).

1.1.1 medizinische Scores

In *Die Intensivmedizin* [Marx et al., 2015] ist der Begriff des Scores folgendermaßen definiert:

> *„Ein Score ist der Versuch, eine komplexe klinische Situation auf einen eindimensionalen Punktwert abzubilden. Eine solche Reduktion verfolgt das Ziel, übergreifende Aspekte wie Schweregrad oder Prognose als Kombination einzelner Fakten objektiv zu fassen, um sie dann in unterschiedlichen Kollektiven vergleichend darstellen zu können.“*

Es handelt sich bei einem Score häufig um die Kombination mehrerer erfassbarer Werte, beispielsweise der Herzfrequenz oder dem Sauerstoffgehalt im Blut. Auch allgemeine Informationen über den Patienten wie das Alter oder bekannte Vorerkrankungen können berücksichtigt werden. Die Bestimmung eines Scores stellt also den Versuch dar, die komplexe, individuelle Situation eines Patienten auf einen numerischen Wert zu reduzieren. Dabei gehen unweigerlich Informationen verloren. Gleichzeitig erlaubt es die Erfassung von derartigen standardisierten Scores aber, auf einen Blick wichtige Informationen über den Zustand des Patienten zu erfassen. Durch eine derartige Reduktion auf das Wesentliche wird ferner ermöglicht, den pathologischen Verlauf eines Patienten über einen längeren Zeitraum zu analysieren, oder die Symptomatik mehrerer Patienten miteinander zu vergleichen. Ein weiterer Vorteil ist es, dass, unter der Voraussetzung der richtigen Anwendung, die Vergabe von Scores weitestgehend unabhängig von der subjektiven Einschätzung des Arztes oder der Pflegekraft erfolgt [Marx et al., 2015].

Die Frage, ob es sich bei der Vorhersage der im Rahmen dieser Arbeit behandelten Scores um ein Regressions- oder ein Klassifikationsproblem handelt, wird in Abschnitt 1.2.1 weiter vertieft. Eine Ausnahme bildet der CAM-ICU (siehe Abschnitt 2.1). Das Ergebnis fällt hierbei entweder positiv oder negativ aus und stellt damit keinen Score im eigentlichen Sinne dar.

1.1.2 Ziel der Arbeit

Das Ziel der vorliegenden Arbeit ist es, mit Hilfe von maschinellem Lernen ein statistisches Modell zu entwickeln, um anhand von Freitexten medizinische Scores möglichst akkurat vorherzusagen.

1.2 Maschinelles Lernen

Der Begriff Maschinelles Lernen bezeichnet einen modernen Ansatz in der Arbeit an künstlicher Intelligenz. (cite).

hat sich im Zeitalter von Big Data und leistungsstarken Rechnern zu einem der Hauptforschungspunkte der Informatik entwicklt (cite).

Mitchell [1997] definiert den Vorgang maschinellen Lernens folgendermaßen:

> *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."*

Im Allgemeinen werden also Algorithmen, die aus großen Datensätzen „lernen“ und damit Vorhersagen über unbekannte Daten machen, als maschinelles Lernen bezeichnet.

Beschreiben: Hier gehts um supervised ML!

1.2.1 Regression vs Klassifikation

Probleme aus dem Bereich des überwachten maschinellen Lernens lassen sich im Allgemeinen in eine von zwei Kategorien einordnen: Klassifizierung bezeichnet den Prozess, bei dem ein Datensatz einer oder mehreren Klassen aus einer endlichen Liste möglicher Klassen zugeordnet wird. Dieser Ansatz findet beispielsweise bei der automatischen Kategorisierung von E-Mails (Spam oder nicht Spam) oder bei der Erkennung von

handschriftlichen Texten (welches Symbol aus einem gegebenen Alphabet ist dargestellt?) Anwendung (CITE). Da es sich bei den betrachteten medizinischen Scores um diskrete, ganzzahlige Werte aus einem endlichen Wertebereich handelt, liegt auch hier die Anwendung eines Klassifizierungs-Verfahrens nahe.

Betrachtet man aber die verschiedenen möglichen Werte eines Scores als separate und voneinander unabhängige Klassen, so ginge eine wichtige Information über deren Anordnung verloren. Bei den im Rahmen dieser Arbeit behandelten Scores handelt es sich stets um eindimensionale, metrische Skalen. Im mathematischen Sinne stellen sie Totalordnungen dar: Sie erfüllen also die Anforderungen der Reflexivität, Antisymmetrie, Transitivität und Totalität. Bezeichne M die Menge aller möglichen Werte eines beliebigen medizinischen Scores. Es gilt also für alle $a, b, c \in M$:

$$
\begin{aligned}
a \leq a & \qquad \text{(Reflexivität)} \\
a \leq b \wedge b \leq a \;\Rightarrow\; a = b & \qquad \text{(Antisymmetrie)} \\
a \leq b \wedge b \leq c \;\Rightarrow\; a \leq c & \qquad \text{(Transitivität)} \\
a \leq b \vee b \leq a & \qquad \text{(Totalität)}
\end{aligned}
$$

Damit lassen sich die verschiedenen Scores vergleichen und in ein Verhältnis setzen. So ist ein RASS-Wert[1] von -4 (tief sediert) beispielsweise deutlich näher an -3 (mäßig sediert) als an $+1$ (unruhig). Bei gängigen Verfahren zur Klassifizierung ginge diese Information verloren, da bei Kenngrößen zur Bewertung solcher Modelle nur betrachtet werden kann, ob ein gegebener Eingabetext genau der richtigen Kategorie (dem richtigen Score) zugeordnet wurde oder nicht.

Bei der vorliegenden Arbeit habe ich mich demnach dafür entschieden, die Vergabe von Scores anhand von Eingabetexten als klassisches Regressionsproblem zu betrachten, und die Ausgaben der Modelle im Zweifelsfall auf den nächstmöglichen ganzzahligen Wert zu runden. Dieser Ansatz fand auch bei früheren Arbeiten, die sich mit ähnlichen Fragestellungen befassten, Anwendung (CITE). Die Abweichung des vorhergesagten Werts eines Modells von dem nächst möglichen Wert stellt somit sogar einen rudimentären Ansatz zur Bewertung der Konfidenz bei der Vergabe einzelner Werte dar.

[1] Richmond Agitation-Sedation Scale

1.2.2 Natural Language Processing

Die Verarbeitung natürlicher Sprache stellt eine besondere Herausforderung im Gebiet des maschinellen Lernens dar.

1.2.3 Anwendungen von Maschinellem Lernen in der Medizin

Lorem Ipsum. Hier andere Einsatzgebiete von ML in Medizin [Krishnan and Sowmya Kamath, 2018]. Und: Warum ML für unsere Problematik?

2 Übersicht über die Daten

information über copra (fridtjof ausquetschen). bla bla bla

Die Daten wurden vor dem Export pseudonymisiert, indem patientenbezogene Daten wie Name und Geburtsdatum entfernt wurden. Insbesondere werden Zeitpunkte nicht in absoluter Form angegebenen, sondern in Relation zum Beginn des Krankenhausaufenthalts des entsprechenden Patienten. Rückschlüsse auf die Identität der Patienten, deren Daten hier betrachtet werden, sind somit ausgeschlossen. Eine Zuordnung der erfassten Werte und Freitexte untereinander zum gleichen Patienten bleibt aber in Form einer eindeutigen Zahlenkombination weiterhin möglich.

Nach dem Export lagen die Daten in Form mehrerer Datendateien vor. `patienten.csv` enthält Metainformationen über die betrachteten Patienten. Separat gibt `delir.csv` für jeden Patienten an, ob für diesen während seines Aufenthalts ein Delir diagnostiziert wurde. Letztendlich geben die beiden Dateien `scores1.csv` und `scores2.xlsx` Aufschluss über die erfassten Scores und eingetragenen Freitexte. Jede Zeile enthält hierbei jeweils die VarID, den betreffenden Patienten, den Zeitpunkt sowie den eigentlichen erfassten Wert. Bei der VarID handelt es sich um einen ganzzahligen Wert, mit der jede Art von auf der Intensivstation erfasstem Wert bzw. eingetragenem Text intern repräsentiert und eindeutig identifiziert wird (siehe Tabelle 2.1).

2.1 Übersicht über die vorliegenden Daten

Der mir für diese Arbeit zur Verfügung stehende Datensatz enthält Informationen über insgesamt 1357 Patienten-Aufenthalte auf den Intensivstationen der Charité Berlin, die allesamt im Zeitraum von XXX 2019 bis XXX 2020 stattfanden. Jeder Aufenthalt wird über eine numerische, inkrementell aufsteigende Nummer[1] eindeutig identifiziert. Hierbei gilt es zu beachten, dass ein Patient bzw. eine Patientin, der/die im Laufe seiner/ihrer Behandlung mehrere Male auf eine Intensivstation verlegt wird, für jeden separaten Aufenthalt eine neue Identifikationsnummer erhält. Für jeden

[1] in den exportierten Datensätzen als n_ID bezeichnet

Patient/für jede Patientin liegen Informationen über das Geschlecht, BMI[2], das Alter zum Zeitpunkt der Aufnahme und ob während des Aufenthalts ein Delir[3] diagnostiziert wurde, vor. Weiterhin ist die Dauer des Aufenthalts auf der Intensivstation vermerkt, sowie ob der Patient/die Patientin während seines/ihres Aufenthalts verstorben ist. Die Mortalität während des Aufenthalts lag bei etwa 17% ($n = 225$). Die beobachteten Aufenthalte verliefen über Zeiträume von wenigen Stunden bis zu mehreren Monaten. Die mittlere Aufenthaltsdauer während des Beobachtungszeitraums liegt bei etwa 7,2 Tagen, der Median beträgt 3,2 Tage. Fast die Hälfte aller Aufenthalte endete also nach weniger als drei Tagen. Nur etwas mehr als ein Viertel der Patienten ($n = 375$) wurden für eine Woche oder länger auf der Intensivstation behandelt.

Für jeden Patienten werden während der Dauer seines Aufenthalts medizinische Scores (siehe Abschnitt 1.1.1) bestimmt sowie Visitentexte geschrieben. Die Visitentexte werden digital erfasst und liegen im Unicode-Zeichensatz vor, folgen allerdings im Allgemeinen keinem einheitlichen Muster. Es handelt sich also um Freitexte, und es liegt im Ermessen der behandelnden Ärzte bzw. Pflegekräfte, einen aussagekräftigen Text zu formulieren. Ebenso können sich Faktoren wie Zeitdruck oder Stress auf Umfang und Genauigkeit der eingetragenen Texte auswirken(CITE). Tabelle 2.1 enthält eine Übersicht über alle Scores und Freitexte, die in dem gegebenen Datensatz erfasst wurden.

Es folgt eine detailliertere Beschreibung derjenigen Werte, die für den Inhalt dieser Arbeit besonders hohe Relevanz haben:

Glasgow Coma Scale Bei der Glasgow Coma Scale (GCS) wird die Schwere einer möglicherweise vorliegenden Bewusstseinsstörung anhand von drei Kategorien (Augenöffnen, verbale Antwort und motorische Reaktion) ermittelt. In jeder Kategorie wird eine Punktzahl ermittelt, die dann zu einem Endergebnis aufsummiert werden. Insgesamt sind so Werte zwischen einschließlich 3 und 15 möglich [Teasdale and Jennett, 1974; Marx et al., 2015].

[2] Body-Mass-Index
[3] F05.* gemäß ICD-10
[4] Entgegen der ursprünglichen Spezifikation ist an der Charité zusätzlich eine Eintragung des Werts 0 (für „wach, voll orientiert“) möglich.

VarID	Name	Wertebereich
20512769	Glasgow Coma Scale (GCS)	$v \in [3, 15]$
20512801	Behavior Pain Scale (BPS)	$v \in [3, 12]$
20512802	Delirium Detection Score (DDS)	$v \in [3, 35]$
22085815	Visite_ZNS	Freitext
22085820	Visite_Oberarzt	Freitext
22085836	Visite_Pflege	Freitext
22085897	Ramsay Sedation Scale	$v \in [0, 6]$[4]
22085911	NRS/VAS (Visual Analogue Scale)	$v \in [0, 10]$
22086067	Vigilanz	Freitext*
22086158	Richmond Agitation Sedation Scale (RASS)	$v \in [-5, 4]$
22086169	CAM-ICU	$v \in \{\text{neg.}, \text{pos.}, \text{unmögl.}\}$
22086170	BPS-Bewertung	Freitext*
22086172	NRS/VAS Bedingungen	Freitext*

Tabelle 2.1: Übersicht aller erfassten Scores und Freitexte

RASS Die Richmond Agitation Sedation Scale (RASS) ist eine zehnstufige Messzahl, die den Grad der Sedierung eines Intensivpatienten beschreibt. Mögliche Werte liegen zwischen -5 („unarousable“/„nicht erweckbar“) und 4 („combative“/„streitlustig“). Bei den betrachteten Patienten wurde der Wert 0 („aufmerksam und ruhig“) am Häufigsten erfasst, was dem erwünschten Wert entspricht, solange keine Sedierung indiziert ist. Die Autoren der Skala empfehlen eine Reihe von Stimuli, die dem Patienten präsentiert werden sollen, um eine besonders genaue Bestimmung des richtigen Wertes zu ermöglichen [Sessler et al., 2002]. Die RASS weist eine hohe Reliabilität und Validität auf und gilt im deutschsprachigen Raum als Goldstandard zum Monitoring der Sedierungstiefe [Marx et al., 2015; Müller et al., 2015].

CAM-ICU bla bla bla

NRS/VAS (Visual Analogue Scale) bla bla bla

Visite_ZNS, Visite_Oberarzt und Visite_Pflege bla bla bla (Fridtjof fragen)

2.1.1 Exemplarische Vorstellung eines Patienten

Aufenthalt des Patienten hier detailliert beschreiben, und seinen Scatterplot einfügen (aber noch ohne Pfeile).

2.1.2 Generierung von Schlüssel-Wert-Paaren

Eine besondere Herausforderung stellte dar, dass die verschiedenen Werte und Texte zu unterschiedlichen Zeiten und unabhängig voneinadner eingetragen werden. Die Informationen über Zeitpunkt und Art der Eintragung sowie der eigentliche Wert liegen jeweils als Tripel in Form mehrerer Log-Dateien vor. Bei den Visitentexten entspricht der eingetragene Text dem Wert der Eintragung.

Um ein Machine-Learning Modell aus dem Bereich des supervised learnings zu trainieren ist eine hohe Anzahl von Trainingspaaren notwendig. Im konkreten Fall dieser Arbeit enthält jedes Wertepaar einen Text sowie einen medizinischen Score, der den in dem Text angegebenen Informationen über die Verfasssung des Patient/die Patientin möglichst genau entspricht. Zusammen bilden diese Paare die Grundlage der Modelle, selbstständig noch nicht gesehene Texte bewerten zu können. Aufgrund der zeitlichen Unterschiede zwischen den Eintragungen von Texten und Scores erwies sich allerdings ebenjene Zuordnung zueinander als nicht trivial.

2.2 Genauigkeit der erfassten Daten

Hier beschreiben, dass viele Texte nicht wirklich den Werten entsprechen. Das mache es schwerer, Modelle zu trainieren, und muss berücksichtigt werden. Gründe:

1. Zeitdruck, Stress bei Ärzten, kann vorkommen dass sie einfach Wert vom letzten mal kopieren
2. Werte können sich innerhalb von Minuten ändern (z.B. RASS von 0 auf -5)
3. Weitere Gründe?

Diese Grüunde sind aber nicht weiter Beobachtungsgegenstand der vorleigenden Arbeit. Dennoch muss das bei der Konzeption, Entwicklung und Bewertung beachtet werden, weil sie ohne weitere Maßnahmen möglicherweise eine obere Schranke für die Performance der Modelle darstellen.

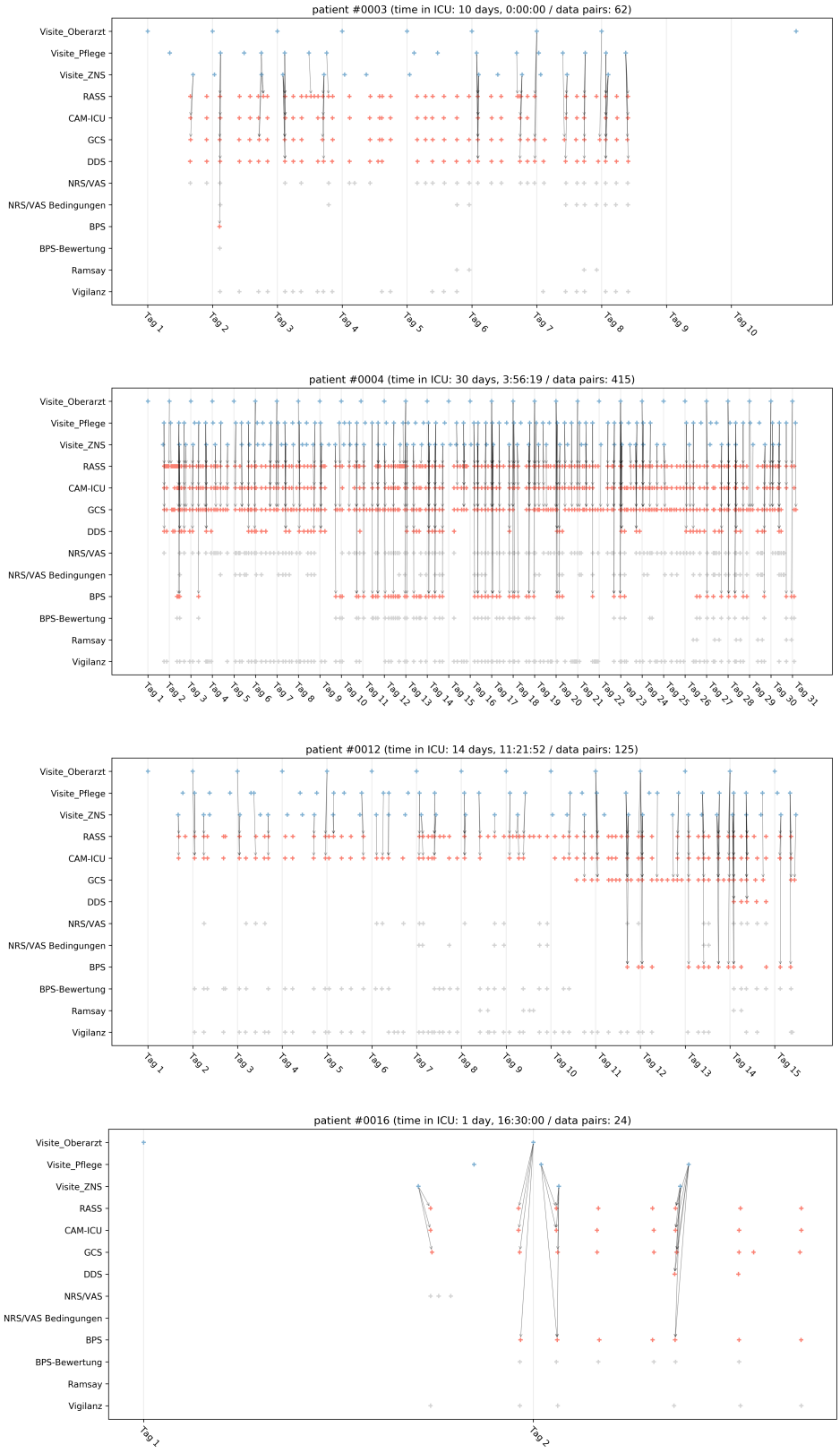

Abbildung 2.1: Übersicht der erfassten Werte einiger Patienten

3 Vorgehensweise

4 Auswertung der Ergebnisse

...

4.1 Bewertung der Leitlinienadhärenz auf den Intensivstationen der Charité

...

4.2 Blick in die Zukunft

...

5 Fazit

Brilliant formuliert, besonders die Konklusio! Langweiliger Alternativtitel: „Fazit“

Literaturverzeichnis

Gokul S. Krishnan and S. Sowmya Kamath. A Supervised Learning Approach for ICU Mortality Prediction Based on Unstructured Electrocardiogram Text Reports. In **Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane**, editors, *Natural Language Processing and Information Systems*, pages 126–134. Springer International Publishing, Cham, 2018. ISBN 978-3-319-91947-8.

Gernot Marx, Elke Muhl, Kai Zacharowski, and Stefan Zeuzem, editors. *Die Intensivmedizin*. Springer Medizin, Berlin Heidelberg, 12., vollständig überarbeitete, akualisierte und erweiterte auflage edition, 2015. ISBN 978-3-642-54952-6 978-3-642-54953-3.

Tom M. Mitchell. *Machine Learning*. McGraw-Hill Series in Computer Science. McGraw-Hill, New York, NY, international ed., [reprint.] edition, 1997. ISBN 978-0-07-115467-3.

Anika Müller, Björn Weiß, Claudia Spies, and S3-Leitliniengruppe. Analgesie, Sedierung und Delirmanagement – Die neue S3-Leitlinie ,,Analgesie, Sedierung und Delirmanagement in der Intensivmedizin“ (DAS-Leitlinie 2015). *AINS - Anästhesiologie · Intensivmedizin · Notfallmedizin · Schmerztherapie*, 50(11/12):698–703, 2015. ISSN 0939-2661, 1439-1074. doi:10.1055/s-0041-107321.

Curtis N. Sessler, Mark S. Gosnell, Mary Jo Grap, Gretchen M. Brophy, Pam V. O'Neal, Kimberly A. Keane, Eljim P. Tesoro, and R. K. Elswick. The Richmond Agitation–Sedation Scale: Validity and Reliability in Adult Intensive Care Unit Patients. *American Journal of Respiratory and Critical Care Medicine*, 166(10):1338–1344, 2002. ISSN 1073-449X, 1535-4970. doi:10.1164/rccm.2107138.

Graham Teasdale and Bryan Jennett. Assessment of Coma and Impaired Consciousness. A Practical Scale. *The Lancet*, 304(7872):81–84, 1974. ISSN 01406736. doi:10.1016/S0140-6736(74)91639-0.

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den 11. Juni 2020