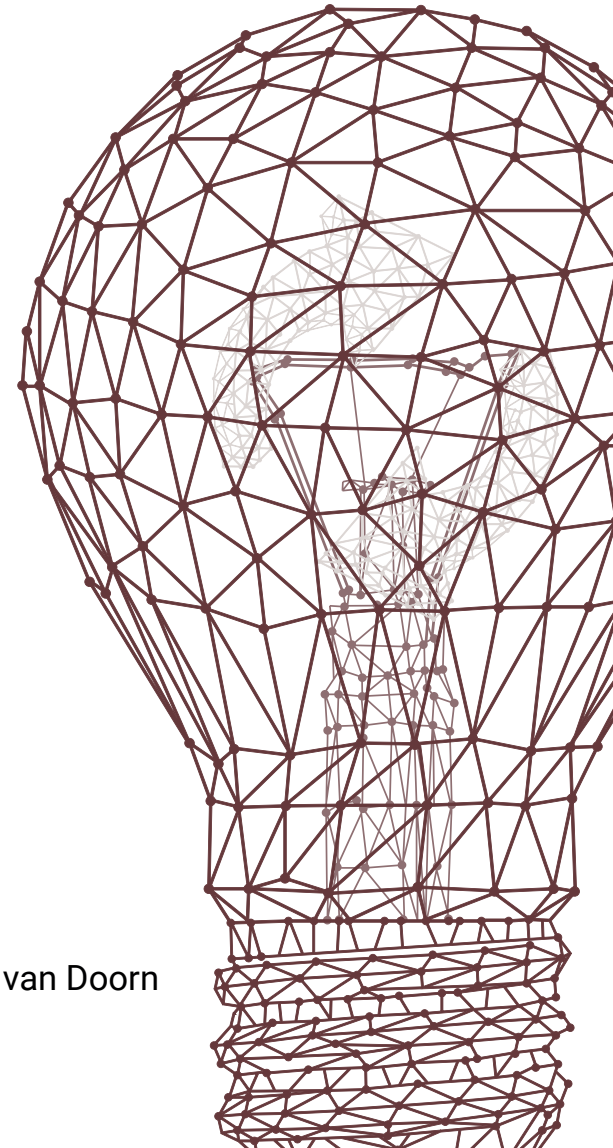
RETHINKING BIOMARKER INNOVATIONS IN LABORATORY MEDICINE



William P.T.M. van Doorn

RETHINKING BIOMARKER INNOVATIONS
IN LABORATORY MEDICINE
William P.T.M. van Doorn

ISBN 978-94-6458-788-3
Layout and cover design by Frank Brouwers
Printed by Ridderprint, Alblasserdam, The Netherlands

RETHINKING BIOMARKER INNOVATIONS IN LABORATORY MEDICINE

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit Maastricht, op gezag van de Rector Magnificus,
Prof. dr. Pamela Habibovic, volgens het besluit van het College van Decanen, in het openbaar te verdedigen
op vrijdag 20 januari 2023 om 13.00 uur

door

William Petrus Theodorus Maria van Doorn

Promotor
Prof. dr. Otto Bekers

Copromotor
Dr. Steven J.R. Meex

Beoordelingscommissie
Prof. dr. Hans-Peter Brunner-La Rocca (voorzitter)
Prof. dr. Rick Body (University of Manchester, Manchester)
Prof. dr. Tilman M. Hackeng
Prof. dr. Ron Kusters (Jeroen Bosch Ziekenhuis, 's Hertogenbosch)
Prof. dr. Leon J. de Windt

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

If you can’t explain it simply,
you don’t understand it well enough.

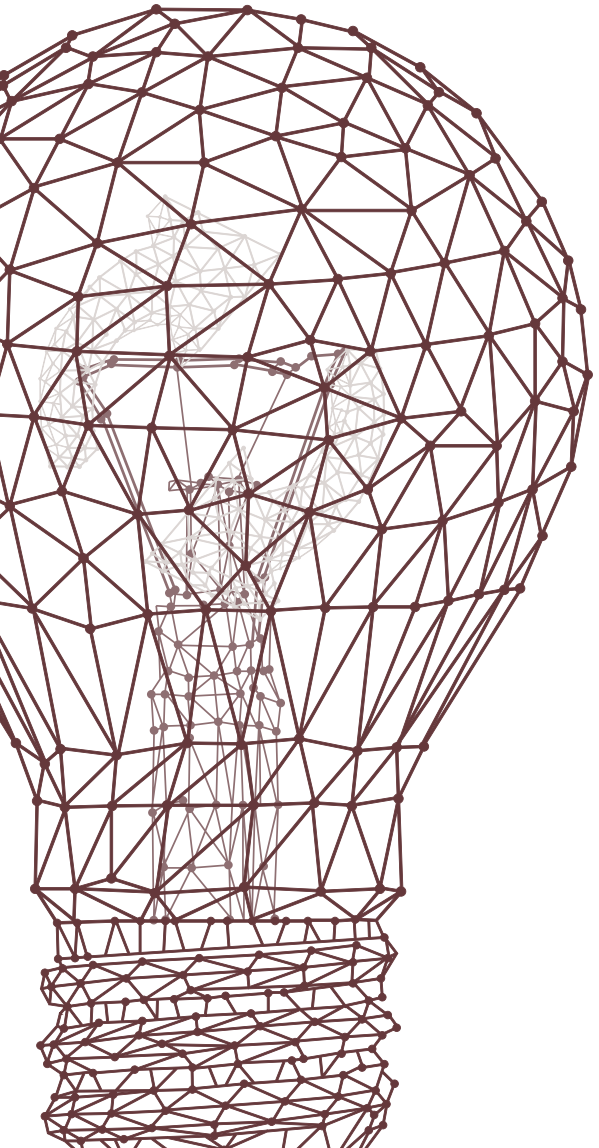
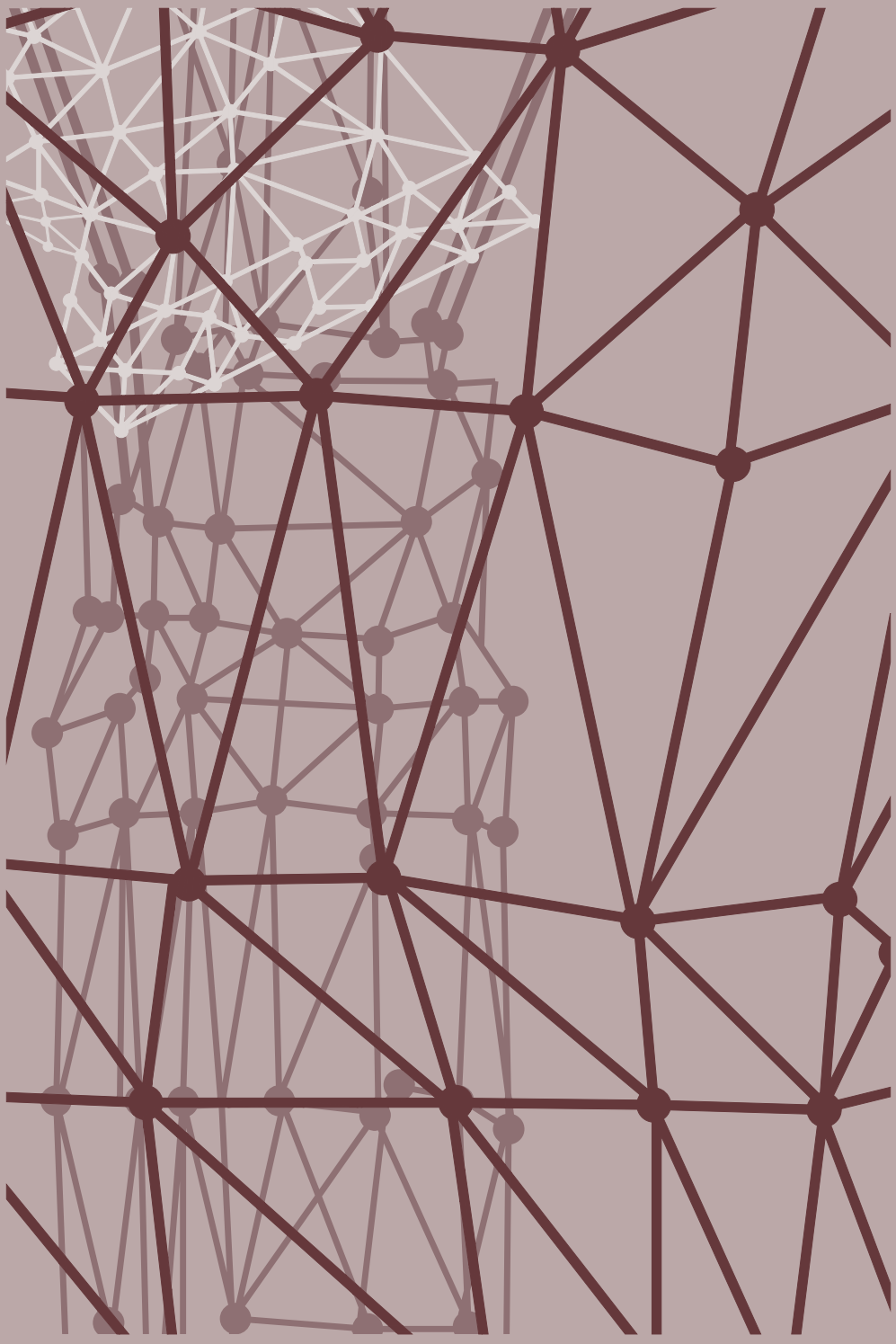


Table of contents



CHAPTER 1

INTRODUCTION

Challenges in the discovery of new biomarkers

The majority of day-to-day clinical decision making is guided by the laboratory measurement of biomarkers[1]. A biomarker is defined as a measurable indicator that reflects the status of a physiological or pathological process, or the response to a medical intervention[2]. A biomarker is expected to enhance the ability of the medical specialist to optimally manage the patient, by confirming or refuting the initial clinical suspicion regarding the patient. Although many biomarkers to date meet this expectation and provide benefit in the diagnostic trajectory, there are many areas in medicine were faster, more accurate, and more sensitive or specific biomarkers would be desirable. Moreover, medical disciplines such as oncology and psychiatry still often lack biomarkers that provide clinical benefit. In other words, there are strong incentives to discover new and better biomarkers. From a research perspective, the biomarker discovery field has substantially evolved in the past decades, empowered by several key developments:

1. Completion of the human genome[3] and HapMap project[4]
2. Development of proteomic-, genomic-, microarray- and nano-technology[5,6]
3. Significant advances in bioinformatics and related disciplines (e.g., computer and data science, system biology and physics)[7]

These advancements accelerated biomarker discovery and development, as reflected by the significant increase of annual publications related to biomarker discovery between 1990 and 2020 (Figure 1).

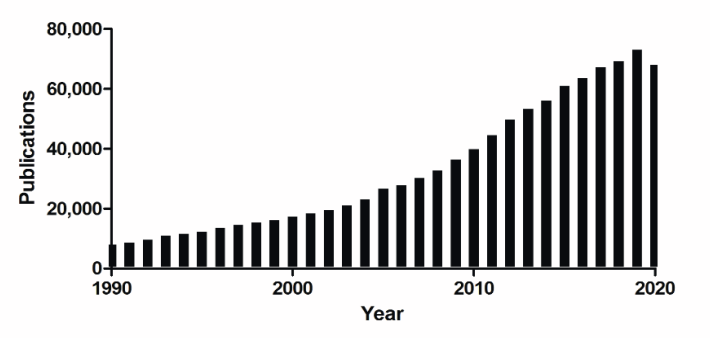


Figure 1: Annual publications related to "biomarker discovery" or "biomarker development" on PubMed®.

Despite these important enablers of biomarker discovery, it is estimated that each year on average less than one diagnostic protein biomarker is approved by the US Food and Drug Administration (FDA)[8]. Thus, a large gap exists between the number of newly discovered biomarkers versus the number of biomarkers actually translating into clinical practice. Reasons for this large disjunction are manifold and reflect the long and difficult path from discovery to clinical use of a biomarker, money, politics, and the lack of coherent and comprehensive pipelines for biomarker development and evaluation[8]. As a means to improve this evidence-based use of biomarkers, researchers argued that a framework for evaluation of new biomarkers should be developed[2,8-10].

A framework for evaluation of existing and new biomarkers

This fostered an ongoing debate in the scientific literature on the most optimal approach for such comprehensive biomarker evaluation, with at least 19 different biomarker evaluation frameworks proposed to date[11]. A systematic evaluation of these frameworks identifies five common features: analytical performance, clinical performance, clinical effectiveness, cost-effectiveness, and broader impact (Table 1)[11,12].

Table 1. Key components in the process of biomarker evaluation.

#	Component	Explanation	Example
1.	Analytical performance	Ability of a laboratory assay to measure biomarker at predefined quality specifications	Universal guidelines recommend that high-sensitivity assays must have an imprecision of ≤10% CV at the 99th percentile of a normal population[13,14].
2	Clinical performance	Ability of a biomarker to confirm to predefined clinical specifications in detecting patients with a particular clinical condition or physiological state	In patients presenting to the emergency department, cardiac myosin-binding protein C (cMyc) was compared with the golden standard (hs-cTn) for diagnosis of myocardial infarction (MI)[15]. At a cut-off of 120 ng/L, cMyc had a 57.1% sensitivity and 96.6% specificity for MI.
3	Clinical effectiveness	Ability of a biomarker to improve health outcomes that are relevant to the individual patient	In patients presenting to the emergency department with acute dyspnea, the clinical effectiveness of B-type natriuretic peptide (BNP) was investigated[16]. Application of BNP reduced length of stay by ~3 days but did not affect 30-day mortality rates.
4	Cost-effectiveness	Ability of a biomarker to be cost-effective compared to the standard situation	A cost-effectiveness analysis examined the introduction of C-reactive protein (CRP) for more appropriate and measured antibiotics use in primary care. The introduction of CRP led to a 12% reduction in costs[17].
5	Broader impact	Consequences of biomarker use beyond elements 1-4, e.g. acceptability, social, physiological, legal, ethical	A review examined the physiological consequences of the introduction of genomics-based noninvasive prenatal tests (NIPT)[18]. NIPT resulted in decreased short-term anxiety and low levels of decisional regret.

This comprehensive framework describes important features for translation of a research biomarker into clinical practice. However, it does not guarantee success, as there is a large body of evidence from biomarkers meeting all requirements yet not being translated into routine clinical care. A telling example is the moderate adaptation of procalcitonin, a biomarker measured in blood that is specific for bacterial infections[19-21]. Procalcitonin has consistently been shown very effective in guiding antibiotic treatment in a wide variety of infections[21-24], hereby performing extremely well according to the five components of biomarker evaluation (Table 1). Even despite procalcitonin outperforming biomarkers that are used for guiding antibiotic treatment (e.g., C-reactive protein[25,26]) in most studies, the real-world implementation of procalcitonin is slow, and failing to reach widespread adoption[27,28]. The opposite also occurs: in fact, many of the current routine biomarkers in clinical chemistry, which were implemented years ago, would not meet the criteria

of biomarker evaluation, if they were discovered today. Even NT-proBNP, a biomarker of which few cardiologists would doubt its effectiveness, failed to reduce hospitalization and cardiovascular mortality in randomized controlled trial where natriuretic peptide-guided therapy was compared to usual care in heart failure patients with reduced rejection fraction[29]. The result of the GUIDE-IT trial should not invalidate or discourage the use of NT-pro BNP –there are a number of plausible arguments why the study may have turned out "false-negative"- but at least it is a telling example of the challenges associated with evidence-based use of new biomarkers.

An alternative approach: optimize current biomarkers rather than deploying new ones

To overcome the challenges associated with deploying a new biomarker in routine clinical care, this thesis pursues an alternative approach to improve the diagnostic trajectory or prognostication of patients by optimizing current biomarkers, rather than deploying new ones. Such approach may benefit from lower costs (compared to traditional development), existing infrastructure, established knowledge in relation to the biomarker biology and detection, and more convenient implementation as the biomarker is likely part of clinical guidelines and thus known by clinicians and laboratory specialists. Four alternative approaches to optimize current routine biomarkers are proposed:

1. improve test precision, accuracy, sensitivity and/or specificity by optimizing the performance characteristics and eliminating interferences of assays that measure these biomarkers (component 1 from Table 1);
2. improve clinical decision limits or diagnostic algorithms that are based upon these biomarkers (component 2);
3. improve clinical specificity by detection of an modified form of these biomarkers, e.g. a post-translational modification or degradation product (component 2 and 3);
4. developing clinical decision support systems that combine multiple biomarker results into a more readily interpretable output for a clinical user, contributing to a better diagnosis or treatment (component 1 to 5). This is an especially interesting novel development that coincides with the rising interest in machine learning and artificial intelligence in medicine.

In this thesis these alternative approaches are examined for cardiac and acute care biomarkers. Approach one is examined by assessing the interference of biotin on the cardiac troponin assay (chapter 3), whereas approach two is also reviewed for the cardiac troponin assay (chapter 2). Approach three is examined by incorporating a specific signature attached to cardiac troponin (chapter 6). Cardiac troponin is released due to cellular damage to cardiomyocytes, regardless

of the underlying mechanism causing the damage. Despite cardiac troponin exhibiting high diagnostic performance for myocardial infarction, it is elevated in other diseases causing cardiomyocyte damage, thereby complicating the diagnosis for myocardial infarction. Hence, the holy grail in the cardiac biomarker research field would be the identification of a biomarker that combines sensitivity for cardiomyocyte cell death (such as troponin) with specificity for hypoxia. As a native molecule, such biomarker is unlikely to exist. Therefore, this thesis proposes to add an innovative signature to cardiac troponin, which allows selective detection of the biomarker of interest when released under conditions of hypoxia, but leaving biomarker release due to non-hypoxic injury triggers undetected. This signature will be based on the detection of 2-nitroimidazole modifications, which are a family of molecules that bind to proteins under conditions of hypoxia[30,31].

Approach four is examined by enhancing the interpretation of biomarker results by combining multiple biomarker results into a readily interpretable output for a clinical user (chapters 7 to 10). Such modelling of complex patterns of biomarkers can be achieved by means of statistical and/or machine learning methods. These methods encompass a set of computational techniques that are theoretically able to condense multiple biomarkers into lower-dimensional and readily interpretable outputs with improved clinical relevance, such as differential myocardial infarction diagnosis or mortality chance within 31 days[32,33].

Biomarkers of cardiovascular disease: cardiac troponin and natriuretic peptides

Several chapters of this thesis focus on biomarkers specifically designed for cardiovascular diseases. Despite the decreasing prevalence of cardiovascular disease in the Western world, myocardial infarction (MI; also known as a heart attack) and heart failure (HF) remain an important burden of morbidity and mortality worldwide[34,35]. In the Netherlands, consisting of approximately 17 million individuals, each year 75.000 and 240.000 people are diagnosed with myocardial infarction and heart failure, respectively36,37. Accurate and rapid discrimination of patients with or without these pathologies facilitates appropriate treatment and prioritization of resources[38-41]. Discussion and review of available biomarkers for MI and HF have been described elsewhere[42-45] and are out of scope for the current thesis. Instead, the focus will on be two biomarkers that are studied in this thesis: cardiac troponin and natriuretic peptides (NP).

The cardiac troponin complex -consisting of the proteins troponin T, I and C- is located within cardiomyocytes and fulfills a key role in cardiomyocyte contraction by regulating the binding between actin and myosin filaments[46]. Cardiac troponin T (cTnT) and I (cTnI) are expressed exclusively in the heart and are therefore considered cardiac specific biomarkers46. Their presence in blood is indicative of myocardial injury. With the release of the first commercial assays measuring cTnT and cTnI in the circulation, these biomarkers quickly became the gold standard for the diagnosis of MI[47,48]. Subsequent generations of

troponin assays led to improvement of their cardiac specificity, analytical sensitivity and assay precision at low concentrations, allowing the diagnosis of MI to be shortened to intervals of 1 to 3 hours[49,50]. Current generation of troponin assays therefore exhibit high diagnostic performance for MI with especially markedly high sensitivity[38,39]. This increase in sensitivity came at the expense of a decreased specificity; elevations of troponin levels in blood turned out not to be specific for myocardial infarction, but were also found to be associated with a variety of other acute and chronic (non-)cardiovascular pathologies[50-52]. It is therefore of particular interest to unravel novel approaches increasing the specificity of the cardiac troponin assay[53,54]. Moreover, increased sensitivity now allows the detection of troponin levels in (apparently) healthy individuals, possibly opening up new applications for prognosis or stratification of individuals with subtle troponin elevations[55,56].

The natriuretic peptide (NP) system is composed of hormones synthesized by the heart, brain and other organs[57,58]. Once released into the circulation, these hormones maintain fluid and pressure homeostasis by modulating cardiac and renal function, thereby potentially influencing myocardial structure and function[59]. To date, four different groups of NPs have been identified including atrial natriuretic peptide (ANP), B-type natriuretic peptide (BNP), C-type natriuretic peptide (CNP) and dendroaspis natriuretic peptide, a D-type natriuretic peptide (DNP), each with its own characteristic functions. In patients with heart failure, natriuretic peptides are secreted in response to the high ventricular filling pressures. This phenomenon was at the basis of application of NPs in acute heart failure (AHF) diagnostics, especially in patients presenting to the emergency department with acute dyspnea[16,60]. Additionally, there is a growing body of evidence showing that various NPs might have utility in a prognostic setting, even in those with asymptomatic or minimally symptomatic signs of heart failure[61], but this still remains controversial[29].

Outline of this thesis

This thesis aimed to investigate the evaluation, optimization and application of cardiovascular and acute care biomarkers.

In the first four chapters cardiac troponin and natriuretic peptides were examined according to the biomarker evaluation framework. Chapter 2 reviews the current clinical laboratory recommendations for high-sensitivity cardiac troponin assays and evaluates differences between Europe and the USA. In Chapter 3 the potential interference of biotin supplements on a cardiac troponin assay in a real-world, cardiac emergency unit population is examined. Next, the kinetics of cardiac troponin in patients suffering from a myocardial infarction are studied in Chapter 4. In Chapter 5 the impact of a potential circadian rhythm of various natriuretic peptides on the diagnosis of acute heart failure is evaluated.

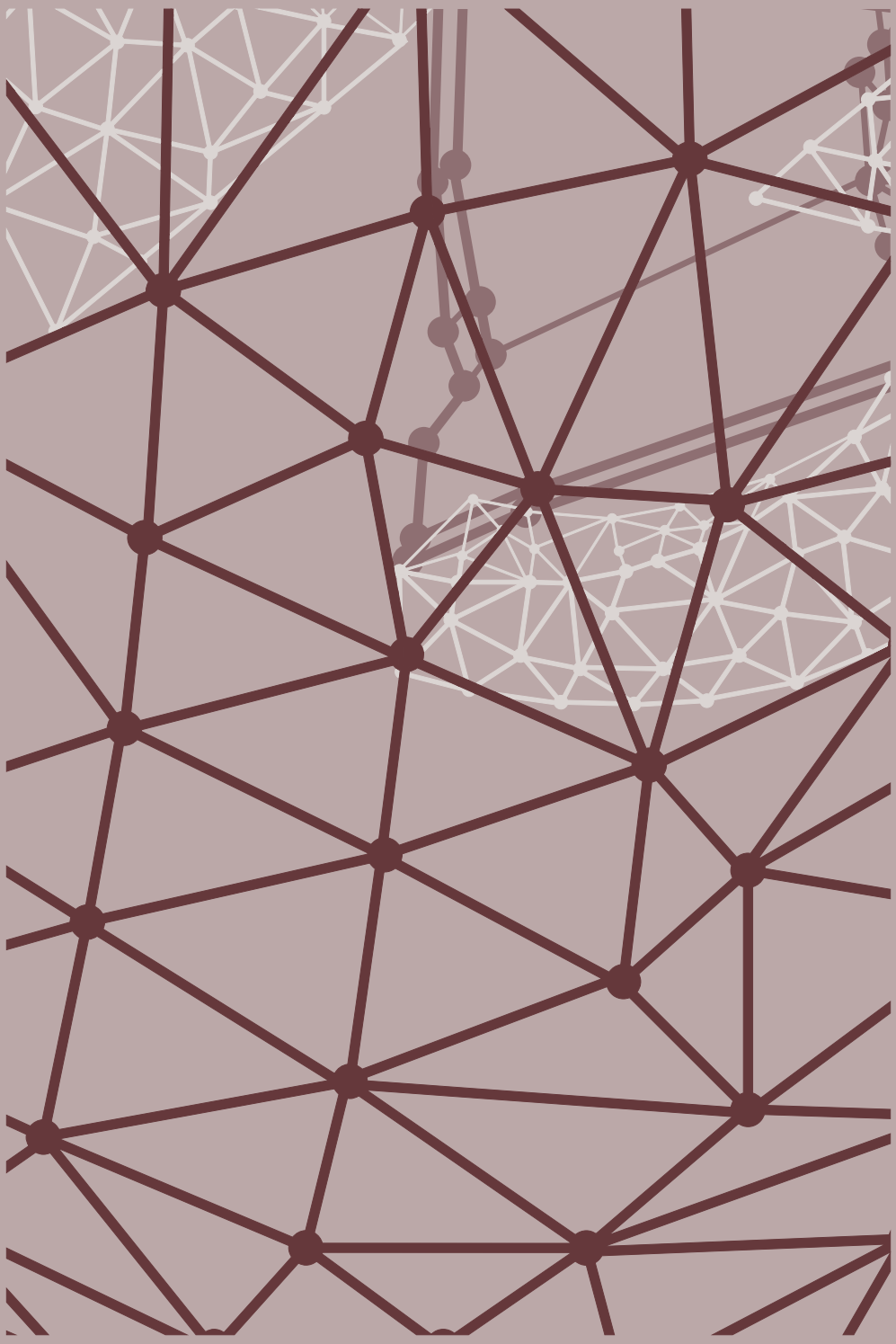
The last five chapters built on alternative approaches to improve the diagnostic trajectory or prognostication of patients by optimizing current biomarkers, rather than deploying new ones. Chapter 6 attempts to improve the clinical specificity of cardiac troponin for the diagnosis of myocardial infarction. As can be deduced from their definition, current "high-sensitive" troponin assays exhibit extremely high sensitivity, but lack a very high specificity. This suboptimal specificity is driven by elevated troponin levels in other (non-) cardiologic pathologies. In an attempt to enhance the specificity of cardiac troponin assays, 2-nitroimidazole molecules were employed enabling the hypoxia-specific detection of cardiac troponin. In Chapters 7 to 10 clinical decision support systems that interpretate multiple biomarker results into a more readily interpretable output for a clinical user were developed. Chapter 7 describes the development of a mortality prediction tool based upon a profile of biomarker levels, and its comparison to the intuition of physicians. The application of this mortality prediction tool was extended in Chapter 8, evaluating its performance in emergency departments of four large hospitals. As a result of the recent COVID-19 pandemic, the risk stratification tool was also validated in this subgroup of patients in Chapter 9. In Chapter 10, historical glucose levels and physical activity data from patients with diabetes were employed to predict future glucose values. The final chapter, Chapter 11, provides a general discussion of the work presented in this thesis and provides directions for future research.

References

1. Manasia, A. & Narimasu, J. in Critical Care (eds John M. Oropello, Stephen M. Pastores, & Vladimir Kvetan) (McGraw-Hill Education).
2. Biomarkers Definitions Working, G. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69, 89-95, doi:10.1067/mcp.2001.113989 (2001).
3. Lander, E. S. et al. Initial sequencing and analysis of the human genome. Nature 409, 860-921, doi:10.1038/35057062 (2001).
4. International HapMap, C. A haplotype map of the human genome. Nature 437, 1299-1320, doi:10.1038/nature04226 (2005).
5. McDermott, J. E. et al. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. Expert Opin Med Diagn 7, 37-51, doi:10.1517/17530059.2012.718329 (2013).
6. Ilyin, S. E., Belkowski, S. M. & Plata-Salaman, C. R. Biomarker discovery and validation: technologies and integrative approaches. Trends Biotechnol 22, 411-416, doi:10.1016/j.tibtech.2004.06.005 (2004).
7. Weston, A. D. & Hood, L. Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. J Proteome Res 3, 179-196, doi:10.1021/pr0499693 (2004).
8. Fuzery, A. K., Levin, J., Chan, M. M. & Chan, D. W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. Clin Proteomics 10, 13, doi:10.1186/1559-0275-10-13 (2013).
9. Horvath, A. R. et al. From biomarkers to medical tests: the changing landscape of test evaluation. Clin Chim Acta 427, 49-57, doi:10.1016/j.cca.2013.09.018 (2014).
10. Steyerberg, E. W. et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 21, 128-138, doi:10.1097/EDE.0b013e3181c30fb2 (2010).
11. Lijmer, J. G., Leeflang, M. & Bossuyt, P. M. Proposals for a phased evaluation of medical tests. Med Decis Making 29, E13-21, doi:10.1177/0272989X09336144 (2009).
12. Bossuyt, P. M., Irwig, L., Craig, J. & Glasziou, P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 332, 1089-1092, doi:10.1136/bmj.332.7549.1089 (2006).
13. Thygesen, K. et al. Fourth universal definition of myocardial infarction (2018). Eur Heart J, doi:10.1093/eurheartj/ehy462 (2018).
14. Giannitsis, E. et al. Analytical validation of a high-sensitivity cardiac troponin T assay. Clin Chem 56, 254-261, doi:10.1373/clinchem.2009.132654 (2010).
15. Kaier, T. E. et al. Direct Comparison of Cardiac Myosin-Binding Protein C With Cardiac Troponins for the Early Diagnosis of Acute Myocardial Infarction. Circulation 136, 1495-1508, doi:10.1161/CIRCULATIONAHA.117.028084 (2017).
16. Mueller, C. et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 350, 647-654, doi:10.1056/NEJMoa031681 (2004).
17. Hunter, R. Cost-effectiveness of point-of-care C-reactive protein tests for respiratory tract infection in primary care in England. Adv Ther 32, 69-85, doi:10.1007/s12325-015-0180-x (2015).
18. Labonte, V., Alsaid, D., Lang, B. & Meerpohl, J. J. Psychological and social consequences of non-invasive prenatal testing (NIPT): a scoping review. BMC Pregnancy Childbirth 19, 385, doi:10.1186/s12884-019-2518-x (2019).
19. Azzini, A. M. et al. A 2020 review on the role of procalcitonin in different clinical settings: an update conducted with the tools of the Evidence Based Laboratory Medicine. Ann Transl Med 8, 610, doi:10.21037/atm-20-1855 (2020).
20. Schneider, H. G. & Lam, Q. T. Procalcitonin for the clinical laboratory: a review. Pathology 39, 383-390, doi:10.1080/00313020701444564 (2007).
21. Assicot, M. et al. High serum procalcitonin concentrations in patients with sepsis and infection. Lancet 341, 515-518, doi:10.1016/0140-6736(93)90277-n (1993).
22. Pepper, D. J. et al. Procalcitonin-Guided Antibiotic Discontinuation and Mortality in Critically Ill Adults: A Systematic Review and Meta-analysis. Chest 155, 1109-1118, doi:10.1016/j.chest.2018.12.029 (2019).

23. Schuetz, P. et al. Effect of procalcitonin-guided antibiotic treatment on mortality in acute respiratory infections: a patient level meta-analysis. Lancet Infect Dis 18, 95-107, doi:10.1016/S1473-3099(17)30592-3 (2018).
24. de Jong, E. et al. Efficacy and safety of procalcitonin guidance in reducing the duration of antibiotic treatment in critically ill patients: a randomised, controlled, open-label trial. Lancet Infect Dis 16, 819-827, doi:10.1016/S1473-3099(16)00053-0 (2016).
25. Tan, M., Lu, Y., Jiang, H. & Zhang, L. The diagnostic accuracy of procalcitonin and C-reactive protein for sepsis: A systematic review and meta-analysis. J Cell Biochem 120, 5852-5859, doi:10.1002/jcb.27870 (2019).
26. Ljungstrom, L. et al. Diagnostic accuracy of procalcitonin, neutrophil-lymphocyte count ratio, C-reactive protein, and lactate in patients with suspected bacterial sepsis. PLoS One 12, e0181704, doi:10.1371/journal.pone.0181704 (2017).
27. Gluck, E. et al. Real-world use of procalcitonin and other biomarkers among sepsis hospitalizations in the United States: A retrospective, observational study. PLoS One 13, e0205924, doi:10.1371/journal.pone.0205924 (2018).
28. Nguyen, C. T., Li, J., Occhipinti, E. A. & Hand, J. Challenges in Procalcitonin Implementation in the Real-World. Open Forum Infect Dis 5, ofy012, doi:10.1093/ofid/ofy012 (2018).
29. Felker, G. M. et al. Effect of Natriuretic Peptide-Guided Therapy on Hospitalization or Cardiovascular Mortality in High-Risk Patients With Heart Failure and Reduced Ejection Fraction: A Randomized Clinical Trial. JAMA 318, 713-720, doi:10.1001/jama.2017.10565 (2017).
30. Krohn, K. A., Link, J. M. & Mason, R. P. Molecular imaging of hypoxia. J Nucl Med 49 Suppl 2, 129S-148S, doi:10.2967/jnumed.107.045914 (2008).
31. Arteel, G. E., Thurman, R. G. & Raleigh, J. A. Reductive metabolism of the hypoxia marker pimonidazole is regulated by oxygen tension independent of the pyridine nucleotide redox state. Eur J Biochem 253, 743-750, doi:10.1046/j.1432-1327.1998.2530743.x (1998).
32. Hastie, T., Tibshirani, R. & Friedman, J. The Elements of Statistical Learning. (Springer New York Inc., 2001).
33. James, G., Witten, D., Hastie, T. & Tibshirani, R. An Introduction to Statistical Learning: with Applications in R. (Springer Publishing Company, Incorporated, 2014).
34. Townsend, N. et al. Cardiovascular disease in Europe: epidemiological update 2016. Eur Heart J 37, 3232-3245, doi:10.1093/eurheartj/ehw334 (2016).
35. Mortality, G. B. D. & Causes of Death, C. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388, 1459-1544, doi:10.1016/S0140-6736(16)31012-1 (2016).
36. Volksgezondheidenzorg.info. Volksgezondheidenzorg.info, <Volksgezondheidenzorg.info> (2016).
37. OECD Better Life Index, 2016).
38. Neumann, J. T. et al. Application of High-Sensitivity Troponin in Suspected Myocardial Infarction. N Engl J Med 380, 2529-2540, doi:10.1056/NEJMoa1803377 (2019).
39. Westwood, M. E. et al. Optimizing the Use of High-Sensitivity Troponin Assays for the Early Rule-out of Myocardial Infarction in Patients Presenting with Chest Pain: A Systematic Review. Clin Chem 67, 237-244, doi:10.1093/clinchem/hvaa280 (2021).
40. Long, B., Koyfman, A. & Gottlieb, M. Diagnosis of Acute Heart Failure in the Emergency Department: An Evidence-Based Review. West J Emerg Med 20, 875-884, doi:10.5811/westjem.2019.9.43732 (2019).
41. Pourafkari, L., Tajlil, A. & Nader, N. D. Biomarkers in diagnosing and treatment of acute heart failure. Biomark Med 13, 1235-1249, doi:10.2217/bmm-2019-0134 (2019).
42. Chan, D. & Ng, L. L. Biomarkers in acute myocardial infarction. BMC Med 8, 34, doi:10.1186/1741-7015-8-34 (2010).
43. Aydin, S., Ugur, K., Aydin, S., Sahin, I. & Yardim, M. Biomarkers in acute myocardial infarction: current perspectives. Vasc Health Risk Manag 15, 1-10, doi:10.2147/VHRM.S166157 (2019).
44. Ibrahim, N. E. & Januzzi, J. L., Jr. Established and Emerging Roles of Biomarkers in Heart Failure. Circ Res 123, 614-629, doi:10.1161/CIRCRESAHA.118.312706 (2018).
45. Spoletini, I., Coats, A. J. S., Senni, M. & Rosano, G. M. C. Monitoring of biomarkers in heart failure. Eur Heart J Suppl 21, M5-M8, doi:10.1093/eurheartj/suz215 (2019).

46. Katrukha, I. A. Human cardiac troponin complex. Structure and functions. Biochemistry (Mosc) 78, 1447-1465, doi:10.1134/S0006297913130063 (2013).
47. Burlina, A., Zaninotto, M., Secchiero, S., Rubin, D. & Accorsi, F. Troponin T as a marker of ischemic myocardial injury. Clin Biochem 27, 113-121, doi:10.1016/0009-9120(94)90021-3 (1994).
48. Alpert, J. S., Thygesen, K., Antman, E. & Bassand, J. P. Myocardial infarction redefined--a consensus document of The Joint European Society of Cardiology/American College of Cardiology Committee for the redefinition of myocardial infarction. J Am Coll Cardiol 36, 959-969, doi:10.1016/s0735-1097(00)00804-4 (2000).
49. Sandoval, Y. et al. 99th Percentile Upper-Reference Limit of Cardiac Troponin and the Diagnosis of Acute Myocardial Infarction. Clinical Chemistry 66, 1167-1180, doi:10.1093/clinchem/hvaa158 (2020).
50. Westermann, D., Neumann, J. T., Sorensen, N. A. & Blankenberg, S. High-sensitivity assays for troponin in patients with cardiac disease. Nat Rev Cardiol 14, 472-483, doi:10.1038/nrcardio.2017.48 (2017).
51. Giannitsis, E. & Katus, H. A. Cardiac troponin level elevations not related to acute coronary syndromes. Nat Rev Cardiol 10, 623-634, doi:10.1038/nrcardio.2013.129 (2013).
52. Smulders, M. W. et al. Initial Imaging-Guided Strategy Versus Routine Care in Patients With Non-ST-Segment Elevation Myocardial Infarction. J Am Coll Cardiol 74, 2466-2477, doi:10.1016/j.jacc.2019.09.027 (2019).
53. deFilippi, C. & Seliger, S. The Cardiac Troponin Renal Disease Diagnostic Conundrum: Past, Present, and Future. Circulation 137, 452-454, doi:10.1161/CIRCULATIONAHA.117.031717 (2018).
54. Mair, J. et al. How is cardiac troponin released from injured myocardium? Eur Heart J Acute Cardiovasc Care, 2048872617748553, doi:10.1177/2048872617748553 (2017).
55. Omland, T. New features of troponin testing in different clinical settings. J Intern Med 268, 207-217, doi:10.1111/j.1365-2796.2010.02253.x (2010).
56. Farmakis, D., Mueller, C. & Apple, F. S. High-sensitivity cardiac troponin assays for cardiovascular risk stratification in the general population. Eur Heart J, doi:10.1093/eurheartj/ehaa083 (2020).
57. Rosenzweig, A. & Seidman, C. E. Atrial natriuretic factor and related peptide hormones. Annu Rev Biochem 60, 229-255, doi:10.1146/annurev.bi.60.070191.001305 (1991).
58. Potter, L. R., Yoder, A. R., Flora, D. R., Antos, L. K. & Dickey, D. M. Natriuretic peptides: their structures, receptors, physiologic functions and therapeutic applications. Handb Exp Pharmacol, 341-366, doi:10.1007/978-3-540-68964-5_15 (2009).
59. Nishikimi, T., Maeda, N. & Matsuoka, H. The role of natriuretic peptides in cardioprotection. Cardiovasc Res 69, 318-328, doi:10.1016/j.cardiores.2005.10.001 (2006).
60. Maisel, A. S. et al. Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure. N Engl J Med 347, 161-167, doi:10.1056/NEJMoa020233 (2002).
61. Oremus, M. et al. BNP and NT-proBNP as prognostic markers in persons with chronic stable heart failure. Heart Fail Rev 19, 471-505, doi:10.1007/s10741-014-9439-6 (2014).



CHAPTER 2

CLINICAL LABORATORY PRACTICE RECOMMENDATIONS FOR HIGH-SENSITIVITY CARDIAC TROPONIN TESTING

William P.T.M. van Doorn*, Wim H.M. Vroemen*, Douwe de Boer,
Alma M.A. Mingels, Otto Bekers, Will K.W.H. Wodzig, Steven J.R. Meex
* equal contribution

JOURNAL OF LABORATORY AND PRECISION MEDICINE
2018;3:30

Introduction

The role of cardiac troponins (cTn) have become increasingly important in diagnosing myocardial infarction (MI), especially in patients without electrocardiogram abnormalities[1]. Since the introduction of high-sensitivity (hs-) cTn immunoassays, there has been extensive clinical guidance on utilizing these biomarkers in patients with acute coronary syndromes[2]. However, recommendations from a laboratory perspective were lacking until the recently reported consensus recommendation from the Academy of the American Association for Clinical Chemistry and the Task Force on Clinical Applications of Cardiac Bio-Markers of the International Federation of Clinical Chemistry and Laboratory Medicine by Wu and colleagues[3]. This globally relevant expert opinion provided ten clinical laboratory practice recommendations associated with hs-cTn testing[3]. These important consensus perspectives were developed to provide global consistency and knowledge in areas where formal guidance and/or data evidence was incomplete. In this Editorial, we not only acknowledge multiple recommendations as defined by Wu and co-workers, but also highlight several specific key aspects from our perspective.

Required hs-cTn guidance for clinicians

Since the launch of the first hs-cTn immunoassay (hs-cTnT, Roche), which was regulatory approved (CE Mark) outside the United States (OUS) in 2010, we recognize that the role of laboratory specialists in educating clinicians, both primary care physicians and clinical specialists, increased significantly. Guidance is predominantly required for patients with a hs-cTn concentration exceeding the MI cut-off threshold, with a rise or fall that is not that obvious, or with a negative coronary angiography. In addition, numerous (pre-)analytical factors and biological variability can lead to cTn results that require guidance from clinical laboratory specialists. Although a lot of research focuses on (patho)physiological properties of cTn and its future potentials, in current clinical practice cTn are biomarkers for MI mandated by guidelines to be evaluated following a serial sampling protocol[4]. As Wu et al. appropriately described, laboratory specialists should educate clinicians on the importance of specific metrics to differentiate clinically relevant hs-cTn concentration changes from analytical and biological variation. As minor hs-cTn changes can have significant clinical impact at a patient level, validating daily quality control (QC), especially at the lower analytical measuring range, is essential. These should preferably be worldwide commutable QC materials for harmonization of the different hs-cTn immunoassays leading to reduction of interassay bias.

hs-cTn cut-off values

The third universal definition of MI recommends cTn testing with a defined cut-off value based on the 99th percentile upper reference limit (URL) of a healthy population[2]. Due to significant sensitivity increases in the most recent generation hs-cTn immunoassays, very low hs-cTn concentrations can be measured with excellent reproducibility (coefficient

of variation (CV) smaller than 10%)[5]. Unfortunately, this significant increase in assay sensitivity led to decreased clinical specificity as detectable hs-cTn concentrations can now be measured in other (non-)pathological conditions in absence of MI[6-8]. In addition to multiple co-morbidities, also age and sex influence hs-cTn concentrations[9, 10]. Consequently, the population used to determine the 99th percentile URL should be carefully selected. Sandoval et al. provided several key recommendations and proposed that multiple surrogate biomarkers should be evaluated to define a healthy population without co-morbidities that influence hs-cTn results[11]. In addition, medical history and medication usage should be taken into account and the population should be diverse with gender, age and ethnicity appropriately distributed[11]. Although we acknowledge their proposal, there is thus far no global consensus on how to define the population used to determine the 99th percentile URL specifically for hs-cTn testing. A perfectly healthy population without hs-cTn influencing co-morbidities and medications will not be a representative population of patients presenting with suspected MI to the emergency department (ED). From our experience, and also reflected by variable MI cut-off values reported in literature, this resulted in a rather heterogeneous implementation of 99th percentiles across clinical laboratories, especially for cTnI[11, 12]. We therefore discourage clinical laboratories to individually determine their own 99th percentile cut-off threshold for MI and recommend them to adapt cut-off values derived from large cohorts in peer-reviewed literature[13-16].

Comparison of hs-cTnI and hs-cTnT

Both hs-cTnT and hs-cTnI provide high diagnostic and prognostic accuracy in patients presenting to the ED with acute chest pain[17]. Therefore, both assays are considered equivalent and laboratories usually implement one hs-cTn immunoassay, which in practice predominantly depends on the clinical chemistry analyzer series used within the clinical laboratory. Nevertheless, it appeared that hs-cTnI seemed to be more prone to outliers compared to hs-cTnT[10, 11, 13]. In addition, harmonization of hs-cTnI assays (currently strictly regulatory cleared OUS; CE Mark) is still an issue due to the heterogeneity of multiple available assays[6]. Apart from analytical heterogeneity, studies conducting hs-cTn assays also highlighted possible biological differences between cTnI and cTnT[17-19]. These include the diurnal rhythm of cTnT versus random fluctuation of cTnI, subtle differences in diagnostic performance and clinical decision limits that are not biologically equivalent for cTnT and cTnI[17-19].

Hs-cTnT assay characteristics

In January 2017, the United States (US) Food and Drug Administration (FDA) cleared the fifth generation cTnT assay by Roche Diagnostics and reported it to be an hs-cTnT assay. Interestingly, the US FDA prescribed assay limits that are not identical to those recommended in OUS CE marked countries. This was mainly due to the fact that different populations were used to determine their respective 99th percentile URL. The limit of

blank (LoB) and limit of detection (LoD), on the other hand, were based on an identical protocol (EP17-A2) and resulted in comparable cut-offs (Table 1)[20]. Additionally, the limit of quantification (LoQ) in the US is 6 ng/L as determined by FDA, while this is 13 ng/L in CE marked countries. This is explained by the fact that the US FDA defined the LoQ at the lowest concentration with a CV ≤ 20% in contrast to a CV ≤ 10% in OUS countries.

Table 1. US and OUS hs-cTnT assay limits as described in the package inserts (Roche Diagnostics).

hs-cTnT assay limits	Module	US hs-cTnT concentration (ng/L)	OUS hs-cTnT concentration (ng/L)
Limit of Blank	e411	3	2.57
	e601/2	2.5	2.26
Limit of Detection	e411	5	4.88
	e601/2	3	2.85
Limit of Quantification	e411	6 (CV 20%)	13 (CV 10%)
	e601/2		

Package insert versions: US; 2018-02, V1.0 – OUS; 2017-03, V9.0.

From a reporting perspective, US clinical laboratories are mandated by the FDA to apply the LoQ (CV ≤ 20%) as the lowest reportable value, while this is less strictly regulated for OUS clinical laboratories. Thus, a very important characteristic for OUS clinical laboratories is to define their lowest reportable hs-cTnT concentration (Table 2). Applying the LoQ would ensure that all reported results are precise, but since serial sampling is advised in European guidelines, we believe that especially a change in hs-cTnT is utterly relevant and therefore recommend to use the LoD (3 ng/L; e601/2) as the lowest reportable hs-cTnT concentration[4].

The importance of blood matrices and cTnT degradation

Solely lithium heparinized (LH) plasma is approved to be used for the hs-cTnT immunoassay in the US while several blood matrices were allowed for the fourth generation immunoassay. Clinical centers in the US should take this into consideration when implementing or transferring to the hs-cTnT assay. Outside the US, multiple blood matrices are allowed but comparing hs-cTnT concentrations across blood matrices is discouraged when applying observation algorithms in suspected MI patients. Recent studies demonstrated altered molecular cTnT form compositions in MI patients between blood matrices with smaller molecules in serum compared to LH plasma[21]. This could lead to altered assay immunoreactivity that potentially influences hs-cTnT results.

In addition to pre-analytical cTnT proteolysis, *in vivo* cTnT fragmentation was also observed in patients suffering from MI and ESRD patients with distinctive molecular compositions[22, 23]. Future research should be performed to investigate the immunoreactivity of these fragments towards the current hs-cTnT assay, but even more importantly, investigate whether specific cTnT fragments could be a target for enhanced assay specificity for MI. This was also recently recognized and suggested by other experts in the field[24, 25].

Table 2. Lowest reportable hs-cTnT concentration scenarios per region and analyzer as described in the package inserts (Roche Diagnostics).

Region	Module	Lowest reportable hs-cTnT concentration (ng/L)
US	e411, e601/2	6 (LoQ at CV 20 %)
OUS	e411	5 (LoD)
		13 (LoQ at CV 10 %)
	e601/2	3 (LoD)
		13 (LoQ at CV 10 %)

Package insert versions: US; 2018-02, V1.0 – OUS; 2017-03, V9.0.

Thus, although the effect of pre-analytical and/or *in vivo* cTnT degradation on the hs-cTnT immunoassay and their direct impact on clinical decisions still remains to be investigated, we recommend OUS clinical laboratories to standardize the blood matrix for hs-cTnT testing. In addition, we advise LH plasma to be used for the most efficient turnaround times promoting clinical decision making. Furthermore, we agree with Wu and colleagues regarding extensive documentation of (pre-)analytical variables when reporting hs-cTn values[3]. This applies both in a clinical and research setting where hs-cTn values are reported.

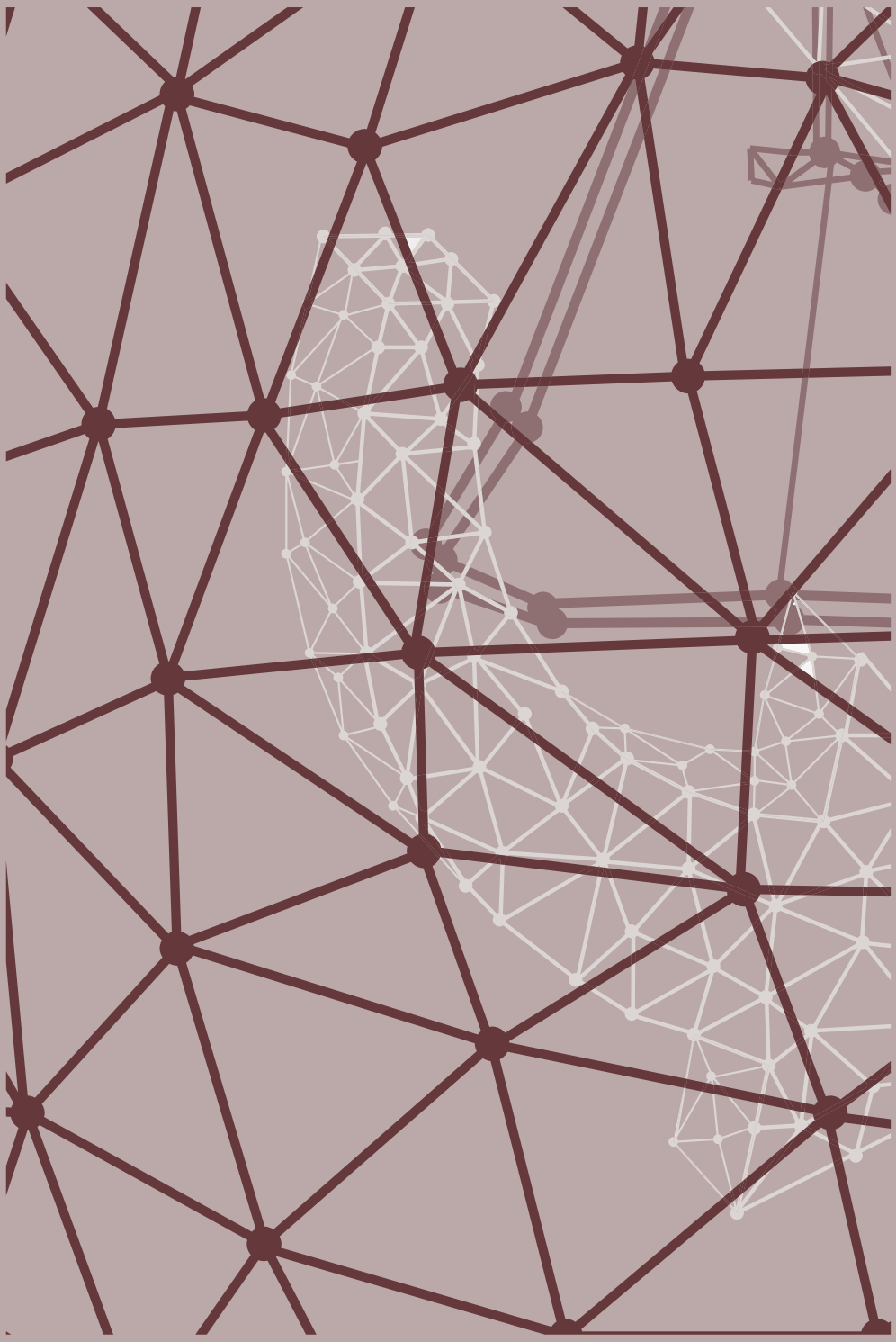
Conclusion

The introduction of hs-cTn immunoassays allowed accurate assessment of hs-cTn concentrations at very low concentrations with excellent precision. Despite its outstanding diagnostic and prognostic value in MI diagnoses, the increase in assay sensitivity led to decreased clinical specificity due to (pre-)analytical and/or (patho-)physiological influences. Guidance, education, and support of clinicians by laboratory specialists will remain essential until hs-cTn specificity for MI is enhanced.

References

1. Hamm CW, Bassand JP, Agewall S, Bax J, Boersma E, Bueno H, Caso P, Dudek D, Gielen S, Huber K, Ohman M, Petrie MC, Sonntag F, Uva MS, Storey RF, Wijns W, Zahger D and Guidelines ESCCfP. ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: The Task Force for the management of acute coronary syndromes (ACS) in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC). Eur Heart J. 2011;32:2999-3054.
2. Thygesen K, Alpert JS, Jaffe AS, Simoons ML, Chaitman BR, White HD, Joint ESCAAHAWHFTFfUDoMI, Authors/Task Force Members C, Thygesen K, Alpert JS, White HD, Biomarker S, Jaffe AS, Katus HA, Apple FS, Lindahl B, Morrow DA, Subcommittee ECG, Chaitman BR, Clemmensen PM, Johanson P, Hod H, Imaging S, Underwood R, Bax JJ, Bonow JJ, Pinto F, Gibbons RJ, Classification S, Fox KA, Atar D, Newby LK, Galvani M, Hamm CW, Intervention S, Uretsky BF, Steg PG, Wijns W, Bassand JP, Menasche P, Ravkilde J, Trials, Registries S, Ohman EM, Antman EM, Wallentin LC, Armstrong PW, Simoons ML, Trials, Registries S, Januzzi JL, Nieminen MS, Gheorghiade M, Filippatos G, Trials, Registries S, Luepker RV, Fortmann SP, Rosamond WD, Levy D, Wood D, Trials, Registries S, Smith SC, Hu D, Lopez-Sendon JL, Robertson RM, Weaver D, Tendera M, Bove AA, Parkhomenko AN, Vasilieva EJ, Mendis S, Guidelines ESCCfP, Bax JJ, Baumgartner H, Ceconi C, Dean V, Deaton C, Fagard R, Funck-Brentano C, Hasdai D, Hoes A, Kirchhof P, Knuuti J, Kolh P, McDonagh T, Moulin C, Popescu BA, Reiner Z, Sechtem U, Sirnes PA, Tendera M, Torbicki A, Vahanian A, Windecker S, Document R, Morais J, Aguiar C, Almahmeed W, Arnar DO, Barili F, Bloch KD, Bolger AF, Botker HE, Bozkurt B, Bugiardini R, Cannon C, de Lemos J, Eberli FR, Escobar E, Hlatky M, James S, Kern KB, Moliterno DJ, Mueller C, Neskovic AN, Pieske BM, Schulman SP, Storey RF, Taubert KA, Vranckx P and Wagner DR. Third universal definition of myocardial infarction. J Am Coll Cardiol. 2012;60:1581-98.
3. Wu AHB, Christenson RH, Greene DN, Jaffe AS, Kavsak PA, Ordonez-Llanos J and Apple FS. Clinical Laboratory Practice Recommendations for the Use of Cardiac Troponin in Acute Coronary Syndrome: Expert Opinion from the Academy of the American Association for Clinical Chemistry and the Task Force on Clinical Applications of Cardiac Bio-Markers of the International Federation of Clinical Chemistry and Laboratory Medicine. Clin Chem. 2018.
4. Roffi M, Patrono C, Collet JP, Mueller C, Valgimigli M, Andreotti F, Bax JJ, Borger MA, Brotons C, Chew DP, Gencer B, Hasenfuss G, Kjeldsen K, Lancellotti P, Landmesser U, Mehilli J, Mukherjee D, Storey RF, Windecker S, Baumgartner H, Gaemperli O, Achenbach S, Agewall S, Badimon L, Baigent C, Bueno H, Bugiardini R, Carerj S, Casselman F, Cuisset T, Erol C, Fitzsimons D, Halle M, Hamm C, Hildick-Smith D, Huber K, Iliodromitis E, James S, Lewis BS, Lip GY, Piepoli MF, Richter D, Rosemann T, Sechtem U, Steg PG, Vrints C, Luis Zamorano J and Management of Acute Coronary Syndromes in Patients Presenting without Persistent STSEotESoC. 2015 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: Task Force for the Management of Acute Coronary Syndromes in Patients Presenting without Persistent ST-Segment Elevation of the European Society of Cardiology (ESC). Eur Heart J. 2016;37:267-315.
5. Giannitsis E, Becker M, Kurz K, Hess G, Zdunek D and Katus HA. High-sensitivity cardiac troponin T for early prediction of evolving non-ST-segment elevation myocardial infarction in patients with suspected acute coronary syndrome and negative troponin results on admission. Clin Chem. 2010;56:642-50.
6. Westermann D, Neumann JT, Sorensen NA and Blankenberg S. High-sensitivity assays for troponin in patients with cardiac disease. Nat Rev Cardiol. 2017;14:472-483.
7. Giannitsis E and Katus HA. Cardiac troponin level elevations not related to acute coronary syndromes. Nat Rev Cardiol. 2013;10:623-34.
8. Gresslien T and Agewall S. Troponin and exercise. Int J Cardiol. 2016;221:609-21.
9. Eggers KM and Lindahl B. Impact of Sex on Cardiac Troponin Concentrations-A Critical Appraisal. Clin Chem. 2017;63:1457-1464.
10. Kimenai DM, Henry RM, van der Kallen CJ, Dagnelie PC, Schram MT, Stehouwer CD, van Suijlen JD, Niens M, Bekers O, Sep SJ, Schaper NC, van Dieijen-Visser MP and Meex SJ. Direct comparison of clinical decision limits for cardiac troponin T and I. Heart. 2016;102:610-6.

11. Sandoval Y and Apple FS. The global need to define normality: the 99th percentile value of cardiac troponin. Clin Chem. 2014;60:455-62.
12. Sandoval Y and Jaffe AS. Using High-Sensitivity Cardiac Troponin T for Acute Cardiac Care. Am J Med. 2017;130:1358-1365 e1.
13. Apple FS, Ler R and Murakami MM. Determination of 19 cardiac troponin I and T assay 99th percentile values from a common presumably healthy population. Clin Chem. 2012;58:1574-81.
14. Gore MO, Seliger SL, Defilippi CR, Nambi V, Christenson RH, Hashim IA, Hoogeveen RC, Ayers CR, Sun W, McGuire DK, Ballantyne CM and de Lemos JA. Age- and sex-dependent upper reference limits for the high-sensitivity cardiac troponin T assay. J Am Coll Cardiol. 2014;63:1441-8.
15. Krintus M, Kozinski M, Boudry P, Capell NE, Koller U, Lackner K, Lefevre G, Lennartz L, Lotz J, Herranz AM, Nybo M, Plebani M, Sandberg MB, Schratzberger W, Shih J, Skadberg O, Chargui AT, Zaninotto M and Sypniewska G. European multicenter analytical evaluation of the Abbott ARCHITECT STAT high sensitive troponin I immunoassay. Clin Chem Lab Med. 2014;52:1657-65.
16. Kimenai DM, Janssen EBNJ, Eggers KM, Lindahl B, den Ruijter HM, Bekers O, Appelman Y and R. MSJ. Sex-specific versus universal clinical decision limits for troponin I and T for the diagnosis of acute myocardial infarction: a systematic review. 2018:Submitted for publication.
17. Rubini Gimenez M, Twerenbold R, Reichlin T, Wildi K, Haaf P, Schaefer M, Zellweger C, Moehring B, Stallone F, Sou SM, Mueller M, Denhaerynck K, Mosimann T, Reiter M, Meller B, Freese M, Stelzig C, Klimmeck I, Voegele J, Hartmann B, Rentsch K, Osswald S and Mueller C. Direct comparison of high-sensitivity-cardiac troponin I vs. T for the early diagnosis of acute myocardial infarction. Eur Heart J. 2014;35:2303-11.
18. Klinkenberg LJ, van Dijk JW, Tan FE, van Loon LJ, van Dieijen-Visser MP and Meex SJ. Circulating cardiac troponin T exhibits a diurnal rhythm. J Am Coll Cardiol. 2014;63:1788-95.
19. Wildi K, Gimenez MR, Twerenbold R, Reichlin T, Jaeger C, Heinzelmann A, Arnold C, Nelles B, Druey S, Haaf P, Hillinger P, Schaerli N, Kreutzinger P, Tanglay Y, Herrmann T, Moreno Weidmann Z, Krivoshei L, Freese M, Stelzig C, Puelacher C, Rentsch K, Osswald S and Mueller C. Misdiagnosis of Myocardial Infarction Related to Limitations of the Current Regulatory Approach to Define Clinical Decision Values for Cardiac Troponin. Circulation. 2015;131:2032-40.
20. Pierson-Perry JF, Format P, Vaks JE and Durham AP. AP17-A2: Evaluation of Detection Capability for Clinical Laboratory Measurement Procedures; Approved Guideline - Second Edition: CLSI; 2012.
21. Katrukha IA, Kogan AE, Vylegzhanina AV, Serebryakova MV, Koshkina EV, Bereznikova AV and Katrukha AG. Thrombin-Mediated Degradation of Human Cardiac Troponin T. Clin Chem. 2017;63:1094-1100.
22. Streng AS, de Boer D, van Doorn WP, Bouwman FG, Mariman EC, Bekers O, van Dieijen-Visser MP and Wodzig WK. Identification and Characterization of Cardiac Troponin T Fragments in Serum of Patients Suffering from Acute Myocardial Infarction. Clin Chem. 2017;63:563-572.
23. Mingels AM, Cardinaels EP, Broers NJ, van Sleeuwen A, Streng AS, van Dieijen-Visser MP, Kooman JP and Bekers O. Cardiac Troponin T: Smaller Molecules in Patients with End-Stage Renal Disease than after Onset of Acute Myocardial Infarction. Clin Chem. 2017;63:683-690.
24. Mair J, Lindahl B, Hammarsten O, Muller C, Giannitsis E, Huber K, Mockel M, Plebani M, Thygesen K, Jaffe AS and European Society of Cardiology Study Group on Biomarkers in Cardiology of the Acute Cardiovascular Care A. How is cardiac troponin released from injured myocardium? Eur Heart J Acute Cardiovasc Care. 2017:2048872617748553.
25. deFilippi C and Seliger S. The Cardiac Troponin Renal Disease Diagnostic Conundrum: Past, Present, and Future. Circulation. 2018;137:452-454.



CHAPTER 3

HIGH-SENSITIVITY CARDIAC TROPONIN I AND T KINETICS AFTER NON-ST-SEGMENT ELEVATION MYOCARDIAL INFARCTION

William P.T.M. van Doorn*, Wim H.M. Vroemen*, Martijn W. Smulders, Jeroen D. van Suijlen, Yvonne J.M. van Cauteren, Sebastiaan C.A.M. Bekkers, Otto Bekers, Steven J.R. Meex
* equal contribution

THE JOURNAL OF APPLIED LABORATORY MEDICINE
2019 Jan;5(1):239-241

Diagnosis of non-ST-segment elevation myocardial infarction (NSTEMI) strongly relies on serial cardiac troponin (cTn) measurements and interpretation of kinetic changes[1]. The current evidence suggests distinct kinetic changes for troponin I and T with a monophasic release curve of cTnI and biphasic, prolonged, release pattern of cTnT[2]. These observations stem from the era of conventional assays, and it is unclear if this is representative for today's high-sensitivity (hs) assays. Another unexplored aspect is the troponin kinetics in patients with NSTEMI. Indeed, previous studies were solely conducted in STEMI[2, 3] and it is conceivable that NSTEMI is characterized by distinct kinetics. The objective of this study was to assess high-sensitivity troponin I and T kinetics in patients with NSTEMI.

Patients who consented to the CARMENTA trial (NCT01559467) were used for this study[4]. Briefly, CARMENTA enrolled patients between 2012 and 2016 who presented to the cardiac emergency unit with clinical symptoms suggestive of NSTEMI, normal or inconclusive electrocardiogram and increased hs-cTnT levels (>14 ng/L at presentation or 3 hours later). The study was approved by the Ethical Committee (11-2-077) and complied with the Declaration of Helsinki. Standardized serum sampling was done at presentation, 3, 6, 24 and 48 hours after presentation. Furthermore, all additional serum samples taken at the discretion of the attending physician for routine clinical care during hospitalization were collected. This analysis included patients with an adjudicated NSTEMI diagnosis of whom ≥3 serum samples were available (n=61). Four patients with a suspected re-infarction during hospitalization were excluded. Troponin concentrations were assessed using the hs-cTnI (Abbott) and hs-cTnT (Roche) immunoassays. The hs-cTnI and hs-cTnT assay have a limit of blank of 0.7-1.3 and 2.26 ng/L, limit of detection of 1.1-1.9 and 2.85 ng/L, and limit of quantification (10% CV) of 4.7 and 5.03 ng/L, respectively. The measuring ranges are 4.7–50,000 ng/L for hs-cTnI and 5.03–10,000 ng/L for hs-cTnT. Time-course of hs-cTnI and hs-cTnT kinetics were modeled using local regression, a non-parametric regression method combining multiple regression models in a k-nearest-neighbors-based meta-model. Data are presented as median[interquartile range]. Statistical analyses were performed with R(3.3.1) and package ggplot2(3.0.0).

The average number of samples per patient was 5.1±1.4. Presenting levels of hs-cTnI were substantially higher than hs-cTnT (173[61-540] ng/L versus 62[26–144] ng/L; p<0.001). These levels increased to peak concentrations of 645[251–2101] ng/L and 157[74-288] ng/L (p<0.001), respectively. Peak concentrations were measured 8.3[6.3–15.7] hours after ED presentation for hs-cTnI and 11.2[6.3–30.9] hours for hs-cTnT (p=0.01). Modeling of hs-cTn kinetic curves revealed a monophasic release pattern for hs-cTnI, whereas a shoulder-like release pattern was observed for hs-cTnT (Figure 1). These patterns were identical when troponin measurements post-revascularization were excluded (data not shown). Pseudo-R^2 of modelled kinetic curves were 0.41 and 0.22 for hs-cTnI and hs-cTnT. Despite substantial unexplained variation in the regression model, the progressively

decreasing elimination rate for hs-cTnT is consistent with the current notion of a biphasic release curve for hs-cTnT[2, 3]. Quantification of the elimination rate showed that hs-cTnI elimination from the circulation was significantly faster than hs-cTnT: 50% reduction of the maximum concentration was observed after 44.6 vs. 115.9 hours (p<0.001).

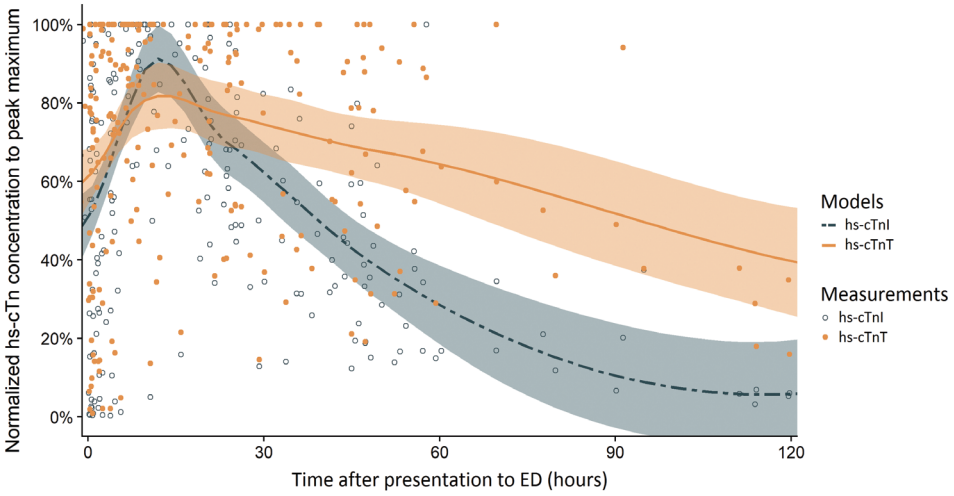


Figure 1. Kinetics of hs-cTnI (dotted line) and hs-cTnT (solid line) modeled in patients with non-ST-segment elevation myocardial infarction after ED presentation. Individual data points are depicted as open dots (hs-cTnI) and closed dots (hs-cTnT). Confidence intervals (95%) are shown as shaded areas surrounding both curves.

This study is the first to model hs-cTnI and hs-cTnT kinetics in NSTEMI patients. We report two major findings.

First, presenting and peak concentrations of hs-cTnI were 3-4 fold higher than hs-cTnT. Since cTnI and cTnT are different molecules, such numerical differences are not surprising. Intriguingly however, the hs-cTnI/hs-cTnT-ratio in these patients is opposite to the ratio observed in patiexnts with chronic hs-cTn elevations, where hs-cTnI is generally lower than hs-cTnT[5].

Second, modeled hs-cTn kinetic curves reveal fast, monophasic elimination of hs-cTnI, and a relatively slow, progressively decreasing elimination rate of hs-cTnT.

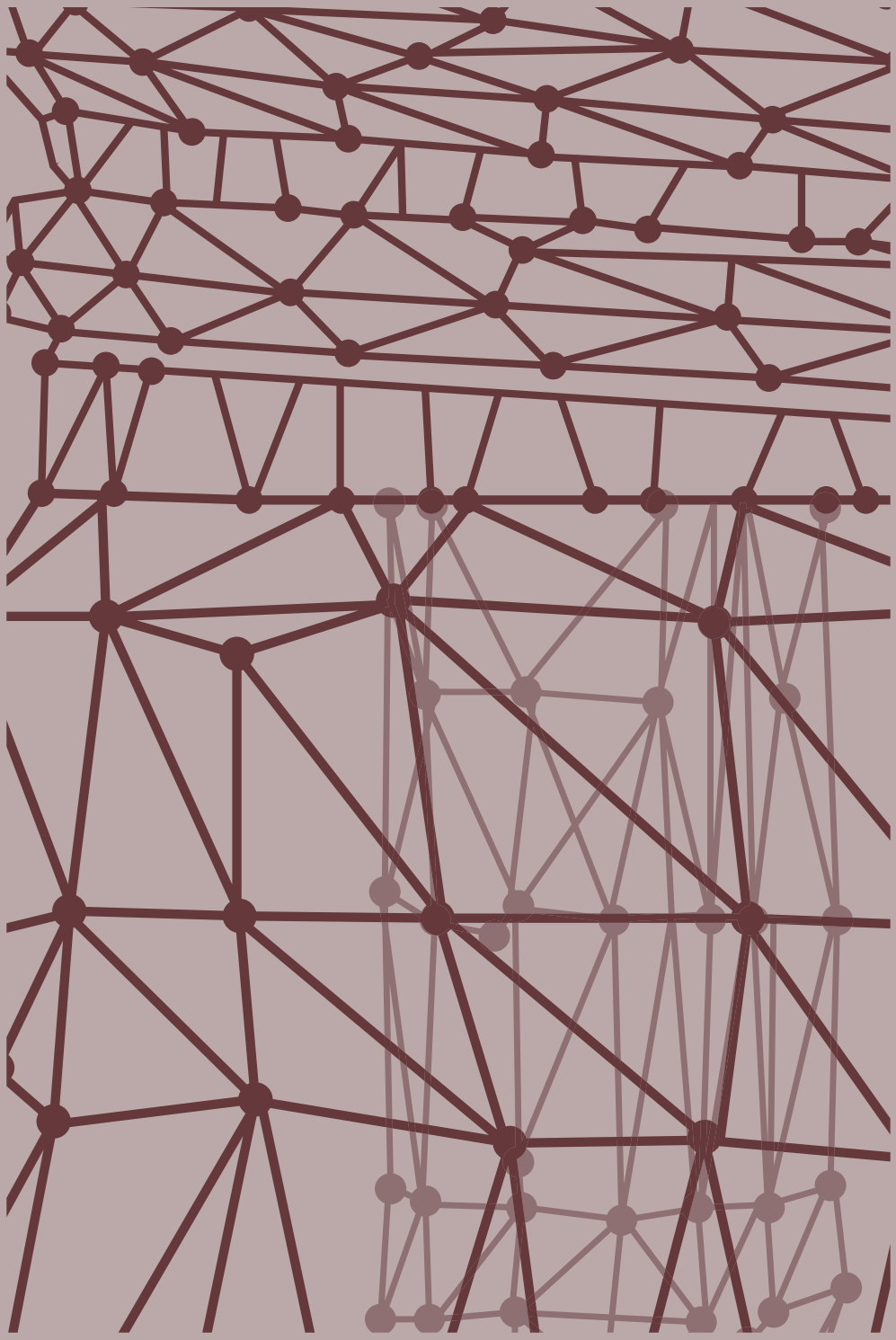
These data confirm and extend the troponin kinetics in patients with STEMI, assessed by conventional and high-sensitivity assays[2,3]. Multiple hypotheses exist for the difference

in kinetics, including cytosolic fraction versus myofibril-bound fraction differences, immunoreactivity, and different kidney clearing pathways(3). However, these are highly suggestive and warrant further investigation. Two limitations of this study merit attention. First, this study is limited as time of pain onset was not obtained. However, our sampling window comprised both the rise and fall of hs-cTnI and hs-cTnT, including peak concentrations. These are the most important reference points to assess kinetic differences of troponin I and T. Second, data density was limited >60 hours. Nevertheless, the increasing confidence intervals do not overlap confirming the difference in kinetics.

In conclusion, using high-sensitivity cardiac troponin assays we show that cTnI and cTnT exhibit distinct kinetics in patients with NSTEMI: fast, monophasic elimination of hs-cTnI, and relatively slow, progressively decreasing elimination of hs-cTnT.

References

1. Thygesen K, Alpert JS, Jaffe AS, Chaitman BR, Bax JJ, Morrow DA, White HD and Group ESCSD. Fourth universal definition of myocardial infarction (2018). *Eur Heart J*. 2019;40:237-269.
2. Katus HA, Remppis A, Scheffold T, Diederich KW and Kuebler W. Intracellular compartmentation of cardiac troponin T and its release kinetics in patients with reperfused and nonreperfused myocardial infarction. *Am J Cardiol*. 1991;67:1360-7.
3. Laugaudin G, Kuster N, Petiton A, Leclercq F, Gervasoni R, Macia JC, Cung TT, Dupuy AM, Solecki K, Lattuca B, Cade S, Cransac F, Cristol JP and Roubille F. Kinetics of high-sensitivity cardiac troponin T and I differ in patients with ST-segment elevation myocardial infarction treated by primary coronary intervention. *Eur Heart J Acute Cardiovasc Care*. 2016;5:354-63.
4. Smulders MW, Kietselaer BL, Das M, Wildberger JE, Crijns HJ, Veenstra LF, Brunner-La Rocca HP, van Dieijen-Visser MP, Mingels AM, Dagnelie PC, Post MJ, Gorgels AP, van Asselt AD, Vogel G, Schalla S, Kim RJ and Bekkers SC. The role of cardiovascular magnetic resonance imaging and computed tomography angiography in suspected non-ST-elevation myocardial infarction patients: design and rationale of the CARdiovascular Magnetic rEsoNance imaging and computed Tomography Angiography (CARMENTA) trial. *Am Heart J*. 2013;166:968-75.
5. Kimenai DM, Martens RJH, Kooman JP, Stehouwer CDA, Tan FES, Schaper NC, Dagnelie PC, Schram MT, van der Kallen CJH, Sep SJS, van Suijlen JDE, Kroon AA, Bekers O, van Dieijen-Visser MP, Henry RMA and Meex SJR. Troponin I and T in relation to cardiac injury detected with electrocardiography in a population-based cohort - The Maastricht Study. *Sci Rep*. 2017;7:6610.



CHAPTER 4

BIOTIN INTERFERENCE IN HIGH-SENSITIVITY CARDIAC TROPONIN T TESTING: A REAL-WORLD EVALUATION IN ACUTE CARDIAC CARE

Wim H.M. Vroemen*, William P.T.M. van Doorn*, Dorien M. Kimenai, Will K.W.H. Wodzig, Douwe de Boer, Otto Bekers, Steven J.R. Meex
*equal contribution

CARDIOVASCULAR RESEARCH
2019 Dec 1;115(14):1950-1951

The United States Food and Drug Administration recently issued a safety communication to warn for biotin interference in cardiac troponin assays.[1] Cardiac troponins are the gold standard biomarkers for diagnosing acute myocardial infarction (AMI). Cardiac troponin concentrations can be falsely low in patients using dietary supplements containing high levels of biotin[1]. Since the prevalence of dietary supplement intake is ≥30% in the United States and Europe, substantial clinical concern has risen that AMI might be missed[1, 2]. Analytical interference of biotin especially applies to cardiac troponin immunoassays exploiting the biotin-streptavidin interaction in the assay configuration[3]. Therefore, we evaluated the real-world prevalence of biotin interference in high-sensitivity cardiac troponin T (hs-cTnT; Roche Diagnostics, Basel, Switzerland) testing in acute cardiac care.

This analysis included 572 consecutive patients of our acute cardiac care unit over a period of 3 months and was carried out according to the principles of the Declaration of Helsinki. No patient informed consent was acquired since the samples used in this analysis were collected as per routine clinical care following the 0/1h algorithm for diagnosing acute myocardial infarction. This biotin interference analysis served as part of an assay verification protocol by our clinical laboratory for routine clinical care. Lithium-heparin plasma samples were collected for routine hs-cTnT concentration assessment according to the 0h/1h diagnostic algorithm for myocardial infarction. The hs-cTnT assay has a limit of blank of 3 ng/L, a limit of detection of 5 ng/L, a limit of quantitation of 13 ng/L, and a linear measuring range of 5 – 10,000 ng/L. To directly assess the effect of biotin-driven hs-cTnT assay interference, hs-cTnT concentrations were assessed before and after biotin depletion using an excessive amount of streptavidin-coated magnetic microparticles. To validate our biotin depletion protocol, biotin concentrations (IDK® Biotin ELISA, Immundiagnostik AG, Bensheim, Germany) were assessed in a subcohort of 100 patients before and after biotin depletion. Considering the applied sample dilution factor (1:2), the biotin ELISA has a limit of blank of 50 ng/L, a limit of detection of 64.8 ng/L, limit of quantitation of 96.2 ng/L and a linear measuring range of 96.2 – 2,200 ng/L.

Median [interquartile range] baseline biotin concentration in the 100 patients subcohort was 331 [219 – 521] ng/L. Of these patients, 11% were biotin deficient (<100 ng/L), 52% had suboptimal concentrations (100 – 400 ng/L) and 37% had optimal concentrations (>400 ng/L)[4]. Our biotin depletion protocol effectively removed almost all free circulating plasma biotin, reducing levels to below the detection limit (96.2 ng/L) in 97% of the samples.

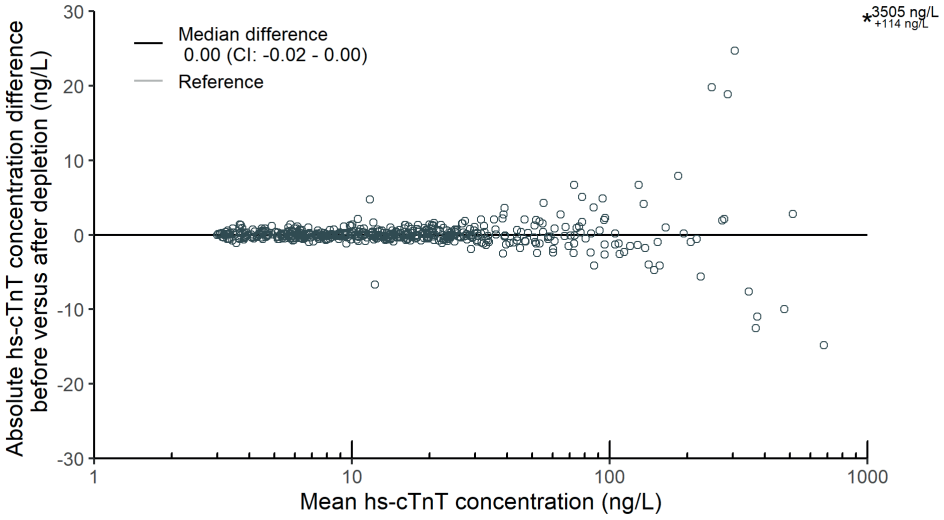


Figure 1. Bland-Altman plot of absolute hs-cTnT concentration differences before and after biotin depletion in 572 patients. The open dots are individual data points. The black line represents the median difference (0.00, 95% CI: -0.02 to 0.00) and the grey line is the reference line.

In the total population, no detectable biotin-associated bias was observed as relative hs-cTnT concentration differences (before minus after biotin depletion) were equally distributed around zero (Figure 1). A Wilcoxon signed-rank test supported unchanged hs-cTnT values after biotin depletion (11.8 [5.6 – 24.2] ng/L versus 11.8 [5.6 – 24.1] ng/L, p = 0.95). Additionally, hs-cTnT differences in all patients were within, or immediately adjacent to, the analytical variation range of the hs-cTnT immunoassay (Figure 1)[5].

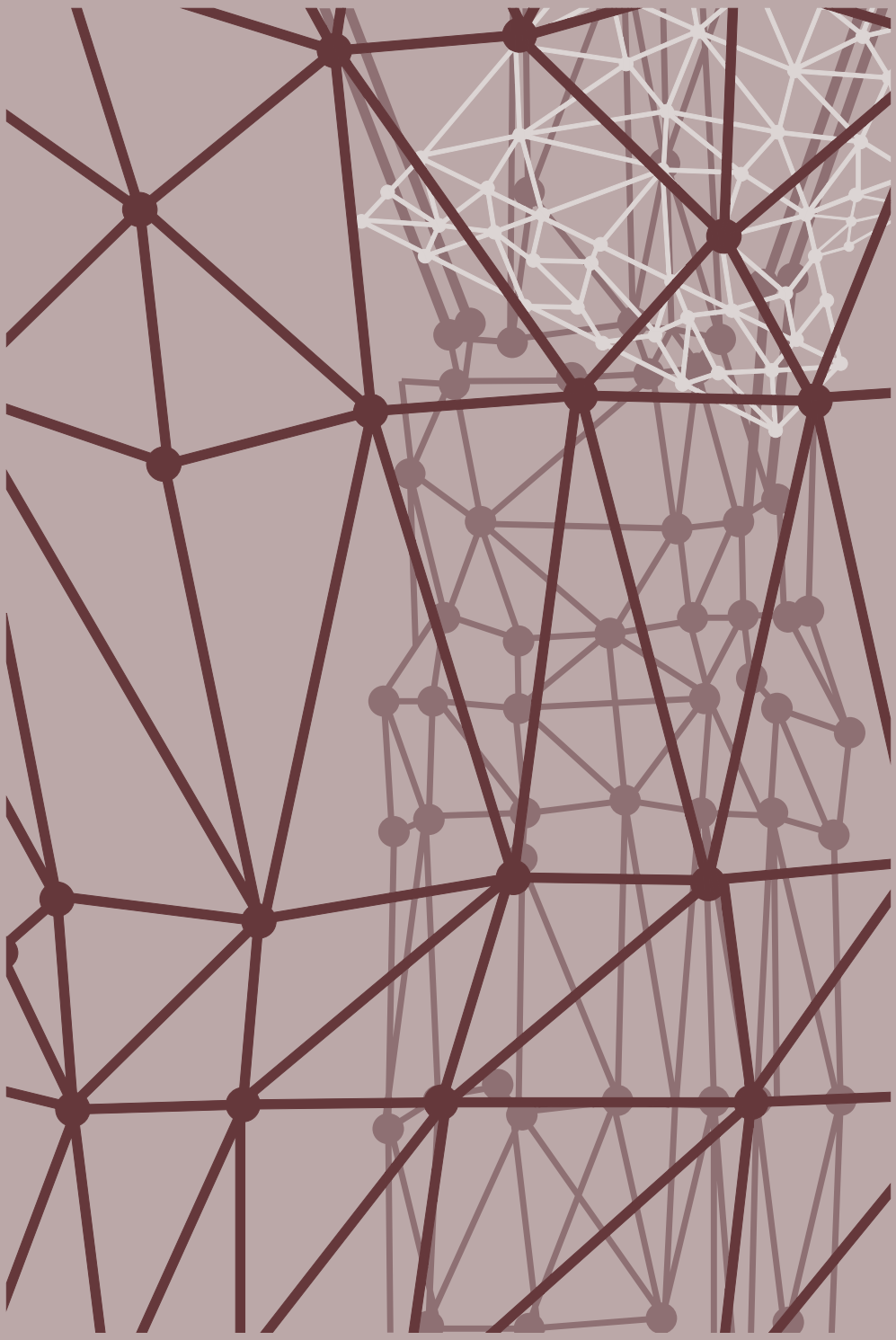
This study is the first to evaluate the real-world prevalence of biotin interference in the hs-cTnT immunoassay in acute cardiac care.
No patient within the subcohort showed biotin levels in the range where assay interference would be suspected, suggesting a very low a priori probability of biotin-driven assay interference in this patient group. In our total acute cardiac care unit population, no single case of biotin interference was found as no relevant change in hs-cTnT concentration after biotin depletion was observed in any patient.

Although biotin's potency at high levels to interfere with various immunoassays is undisputed from an analytical perspective, the present analysis shows that biotin interference is in fact rare. In terms of absolute risk, the probability of a missed AMI diagnosis due to biotin interference is lower than other relatively common sources of risks such as blood sample hemolysis, heterophilic antibodies, patient/blood sample misidentification, or even biological variation of cardiac troponin T[6].

Two limitations of our study merit attention. First, dietary supplement intake information was not collected. However, considering the number of patients included in this analysis and, a dietary supplement intake prevalence of ≥30% in The Netherlands, a substantial population prone for biotin interference in hs-cTnT testing was studied. Second, specific patient groups may receive extremely high-dosages of biotin, e.g. patients with multiple sclerosis and other inflammatory diseases[3]. Even though the risk of a missed AMI diagnosis at a population level may be low, the risk in these specific patient groups is conceivable. In light of these limitations, the anticipated hs-cTnT immunoassay adjustments to abolish the risk of biotin interference is an important improvement to further minimize risk and maximize the diagnostic accuracy of this pivotal AMI biomarker.

References

1. U.S. Food and Drug Administration. The FDA warns that biotin may interfere with lab tests: FDA safety communication. November 28, 2017 (Accessed 29 May 2019). https://www.fda.gov/MedicalDevices/Safety/AlertsandNotices/ucm586505.htm.
2. Pajor EM, Eggers SM, Curfs KCJ, Oenema A and de Vries H. Why do Dutch people use dietary supplements? Exploring the role of socio-cognitive and psychosocial determinants. Appetite. 2017;114:161-168.
3. Saenger AK, Jaffe AS, Body R, Collinson PO, Kavsak PA, Lam CSP, Lefevre G, Omland T, Ordonez-Llanos J, Pulkki K and Apple FS. Cardiac troponin and natriuretic peptide analytical interferences from hemolysis and biotin: educational aids from the IFCC Committee on Cardiac Biomarkers (IFCC C-CB). Clin Chem Lab Med. 2019;57:633-640.
4. Trueb RM. Serum Biotin Levels in Women Complaining of Hair Loss. Int J Trichology. 2016;8:73-7.
5. Giannitsis E, Kurz K, Hallermayer K, Jarausch J, Jaffe AS and Katus HA. Analytical validation of a high-sensitivity cardiac troponin T assay. Clin Chem. 2010;56:254-61.
6. Herman DS, Kavsak PA and Greene DN. Variability and Error in Cardiac Troponin Testing: An ACLPS Critical Review. Am J Clin Pathol. 2017;148:281-295.



CHAPTER 5

DIURNAL VARIATIONS IN NATRIURETIC PEPTIDE LEVELS: CLINICAL IMPLICATIONS FOR THE DIAGNOSIS OF ACUTE HEART FAILURE

Tobias Breidthardt*, William P.T.M. van Doorn*, Noreen van der Linden, Matthias Diebold, Desiree Wussler, Isabelle Danier, Tobias Zimmermann, Samyut Shrestha, Nikola Kozhuharov, Maria Belkin, Caroline Porta, Ivo Strebel, Eleni Michou, Danielle M. Gualandro, Albina Nowak, Steven J.R. Meex, Christian Mueller
*equal contribution

CIRCULATION HEART FAILURE
2022;15(6):e0z09165

Abstract

Background: Current guidelines recommend interpreting concentrations of natriuretic peptides (NPs) irrespective of the time of presentation to the emergency department (ED). We hypothesized that diurnal variations in NP-concentration may affect their diagnostic accuracy for acute heart failure (AHF).

Methods: In a multicenter diagnostic study enrolling patients presenting with acute dyspnoea to the ED and using central adjudication of the final diagnosis by two independent cardiologists, the diagnostic accuracy for AHF of B-type natriuretic peptide (BNP), N-terminal pro-BNP (NT-proBNP), and midregional pro-ANP (MR-proANP) was compared among 1577 "day-time"-presenters versus 908 "evening/night-time"-presenters. In a validation study, the presence of a diurnal rhythm in BNP- and NT-proBNP-concentrations was examined by hourly measurements in 44 stable individuals.

Results: Among patients adjudicated to have AHF, BNP, NT-proBNP and MR-proANP concentrations were comparable among day-time versus evening/night-presenters (all p=ns). Contrastingly, among patients adjudicated to have other causes of dyspnoea, evening/night-presenters had lower BNP (median 44ng/L[18-110] versus 74ng/L[27-168],p<0.01) and NT-proBNP (median 212ng/L[72-581] versus 297ng/L[102-902], p<0.01) concentrations versus day-time-presenters. This resulted in higher diagnostic accuracy as quantified by the area under the curve (AUC) of BNP and NT-proBNP among evening/night presenters [0.97 (95%CI 0.95-0.98) and 0.95 (95%CI 0.93-0.96) versus 0.94 (95%CI 0.92-0.95) and 0.91 (95%CI 0.90-0.93)] among day-time presenters (both p<0.01). These differences were not observed for MR-proANP. Diurnal variation of BNP and NT-proBNP with lower evening/night concentration was confirmed in 44 stable individuals (p<0.01).

Conclusion: BNP and NT-proBNP, but not MR-proANP, exhibit a diurnal rhythm that results in even higher diagnostic accuracy in evening/night-presenters versus day-time-presenters.

Introduction

Acute heart failure (AHF) is the most common cause of unplanned hospitalization, and associated with high morbidity and mortality[1-3] The clinical introduction of natriuretic peptides (NPs) as quantitative markers of hemodynamic stress and heart failure has substantially improved the rapid detection and/or rule-out of AHF among patients presenting with acute dyspnoea[1-4]. Current European and American clinical practice guidelines endorse the diagnostic use of NPs with a class I recommendation [5, 6]. They also suggest interpreting NP-concentrations and applying NP cut-off levels irrespective of the time of presentation. This strategy has recently been challenged by pilot studies suggesting NP-concentrations to display naturally occurring diurnal variations in healthy individuals and in stable patients, with lower NP-concentrations in the evening/night versus during the day [7-10]. It is currently unknown, whether these diurnal variations also occur among patients presenting with acute dyspnoea and thereby possibly affect their diagnostic accuracy.

We therefore aimed to quantify diurnal variations and the possible effect on their diagnostic accuracy for AHF in B-type natriuretic peptide (BNP), N-terminal proBNP (NT-proBNP) and mid-regional pro-atrial natriuretic peptide (MR-proANP) concentrations among day-time versus evening/night-time presenters in a large, multicenter diagnostic study using central adjudication. To validate our findings, we examined diurnal variation with hourly blood sampling over one full day (25 hours) in 44 stable volunteers with or without chronic kidney disease.

Methods

BASEL V ED Patient Population

Basics in Acute Shortness of Breath Evaluation Study (BASEL V) (NCT01831115) was a prospective, diagnostic, multi-centre study enrolling adult patients presenting to the ED of two University (Basel and Zurich) hospitals in Switzerland with nontraumatic acute dyspnoea[11-14]. While enrollment was independent of renal function, patients with terminal renal failure on chronic renal replacement therapy were excluded. For this analysis patients were eligible if the time of presentation was recorded and at least one NP-measurement was performed at presentation (Supplemental Figure 1). The study was carried out according to the principles of the Declaration of Helsinki and approved by the local ethics committees. The authors designed the study, gathered, analyzed and report the data according to the STARD guidelines for studies of diagnostic accuracy.

Central adjudication of AHF (reference standard)

Two independent cardiologists centrally adjudicated the final diagnosis mainly responsible for acute dyspnoea using all medical information pertaining to the patient including detailed clinical assessment, NPs, cardiac imaging, response to therapy, autopsy data for deceased patients, and 90-day follow-up according to the European Society of Cardiology clinical practice guidelines. In situations of diagnostic disagreement, cases were reviewed and adjudicated in conjunction with a third cardiologist.

Biomarker Sampling

At presentation to the ED, blood samples for determination of NP-concentrations were collected into tubes containing potassium EDTA. After centrifugation, samples were either immediately analysed (the NP in clinical use at the respective site) or frozen at −80 °C for later analysis in batch (the other NPs).

Maastricht Serial Sampling Patient Population

Hourly blood sampling over one full day (25 hours) was performed in 44 individuals divided into two study groups as previously described [15-17]. Briefly, the first study group consisted of 24 individuals without clinically diagnosed chronic kidney disease (CKD, 21% females), and the second group consisted of 20 patients (30% females) with clinically diagnosed CKD stage 3 or higher (eGFR, <60 mL/min/1.73 m^2). The estimated glomerular filtration rate (eGFR) was calculated according to the CKD Epidemiology Collaboration formula [18]. Exclusion criteria included current dialysis treatment, myocardial infarction in the 12 months before the study, active cardiac disease (cardiomyopathy, angina pectoris, or myocarditis), and anemia (hemoglobin less than 10.5 g/dL) [15]. Subjects arrived at the laboratory by public transport or car after an overnight fast. During 25h, from 8.30 A.M. till 9.30 A.M. the next day, subjects were restricted to the laboratory environment, and samples were collected every hour from an antecubital venous catheter. Extension lines

for blood sampling were used to prevent disturbance of participants' sleep during the night. Meals were consumed at 8:30 A.M., 12:30 P.M. and 6.00 P.M. (breakfast, lunch and dinner, respectively). Subjects went to bed at 11.30 P.M. and lights were off between 11.35 P.M. and 7.00 A.M. Participants were asked to refrain from exhaustive physical activities and exercise training, two days before the test day. Hemoglobin and hematocrit values were used to quantify possible plasma volume changes due to changes in hydration status and/or posture during the diurnal variation study. The serial sampling study was approved by the Institutional Review Board and Ethics Committee of Maastricht University Medical Center, and registered at clinicaltrial.gov (NCT02091427 and NCT02210897). All participants provided written informed consent. This analysis includes 21 and 19 patients with and without CKD, respectively, after exclusion of 4 individuals due to absence of both BNP and NT-proBNP measurements.

Measurement of NPs

For both cohorts BNP was measured by a microparticle enzyme immunoassay (AxSym or Architect; Abbott Laboratories, Abbott Park, IL, USA) which had a limit of Detection (LoD) of 10 ng/L and an assay range of 10 to 5000 ng/L (package insert). NT-proBNP levels were determined by a quantitative electrochemiluminescence immunoassay (Elecsys proBNP; Roche Diagnostics AG, Zug, Switzerland). According to the package insert, this method has a LoD of 5 ng/L and an assay range of 5 to 35000 ng/L. In the BASEL V ED patient population MR-proANP was measured with an automated sandwich chemiluminescence immunoassay on the KRYPTOR System (BRAHMS AG, Hennigsdorf/Berlin, Germany) as described previously [19]. The functional assay sensitivity (interassay coefficient of variance <20%) is 20 pmol/L.

Statistical analysis

Discrete variables are expressed as counts (percentage) and continuous variables as means ± standard deviation (SD) or median and interquartile range [IR], unless stated otherwise. Comparisons between groups were made using independent Student's t-test and ANOVA, or Mann-Whitney test and Kruskal-Wallis test, Wilcoxon signed rank test, and Pearson's X^2 test, as appropriate. The Jonckheere-Terpstra trend test was used to assess time dependent trends in NP levels. Receiver-operating-characteristic (ROC) curves were constructed to assess sensitivity, specificity and 95% confidence intervals of NP levels to diagnose AHF. The comparison of areas under independent ROC curves (AUC) was performed as recommended by Hanley et al.[20]. For diurnal variation analyses in the BASEL V Emergency Department cohort, patients were separated into quintiles according to their ED presentation time. Patients presenting during the first three presentation time quintiles (9:30 to 15:40) were defined as "day-time" presenters. Patients presenting during the last two presentation time quintiles (15:41 to 9:29) were defined as "evening/night-time" presenters.

Diurnal rhythms of NPs in the Maastricht serial sampling cohort were analyzed by fitting the data to a cosine curve by using the method of cosinor rhythmometry which was extensively described previously [15-17, 21, 22]. Briefly, the cosinor model is described with $Z(t) = M + A \cdot \cos(\omega t + \varphi) + e(t)$; where Z(t) represents the measured natriuretic peptide concentration at a given time (t), M the mesor (value around which oscillation occurs), A the amplitude (half the difference between the peak and the nadir value), ω the angular frequency (degrees per unit time with 360° representing a complete cycle), φ the acrophase (timing of maximal value in degrees), and e(t) the error between the cosinor model and the measurement. Evidence of a diurnal rhythm was indicated by a significant cosinor model fit. Cosine curves on group level were expressed as deviation (%) from the 24h mesor concentration. For both cohorts all hypothesis testing was two-tailed, and p-values less than 0.05 were considered statistically significant unless otherwise stated. Secondary analyses were performed in both cohorts assessing the impact of body weight on NP rhythm. In line with previous studies, patients with a body mass index between 18 and 25 kg/m^2 were considered lean, patients with a body mass index above 30 kg/m^2 were considered obese[10]. Statistical analyses were performed with SPSS for Windows 23.0 (IBM, Armonk, NY), MedCalc 11.2.1.0. (MedCalc Software, Ostend, Belgium), R (version 3.3.1) and the R-packages Cosinor (version 1.1) and Cosinor2 (version 0.1) [22].

Results

BASEL V ED Patient Population: Baseline Characteristics

Among 2485 eligible patients, median age was 76 years, and 43% of patients were women (Table 1A/B). Overall, baseline characteristics were comparable in patients presenting during day-time versus in the evening/night. AHF was the adjudicated final diagnosis in a similar proportion of patients presenting during day-time versus in the evening/night (59% vs. 58%).

Table 1A: Baseline Characteristics

	All patients (n=2485)	Day-time (n=1577)	Evening/Night-time (n=908)
Demographics			
- Age (yrs)	76 [63-83]	76 [64-83]	74 [61-83]
- Male (%)	1415 (57%)	886 (56%)	529 (58%)
Medical history			
- Hypertension (%)	1736 (70%)	1104 (70%)	632 (70%)
- Diabetes (%)	590 (24%)	366 (23%)	224 (25%)
- Coronary Artery Disease (%)	927 (37%)	580 (37%)	347 (38%)
- CKD (%)	802 (32%)	513 (33%)	289 (32%)
- COPD (%)	775 (31%)	474 (30%)	301 (33%)
Physical exam at ED			
- sBP (mmHg)	139±26	136±27	141±26
- HR (beats/min)	91±24	91±23	92±24
- Temperature (°C)	37.1±0.9	37.1±0.9	37.1±0.8
- Oxygen saturation (%)	95±5	94±5	94±5
- Body weight (kg)	74 [63-87]	74 [63-86]	76 [64-87]
- Body Mass Index (kg/m^2)	26 [23-30]	26 [23-30]	26 [23-30]
- Peripheral edema (%)	1118 (45%)	733 (47%)	385 (43%)
- Rales (%)	1168 (47%)	730 (47%)	430 (49%)
Laboratory parameters			
- Hemoglobin (g/L)	132 [117-145]	133 [119-147]	131 [116-144]
- White Blood Count (g/L)	8.8 [7.0-11.5]	8.8 [7.1-11.5]	9.1 [7.0-12.1]
- C-reactive protein (mg/L)	11.6 [3.6-38.0]	12.9 [4.2-40.5]	9.8 [3.3-36.8]
- Creatinine (μmol/L)	89 [70-124]	91 [70-127]	90 [71-123]
- eGFR (ml/min)	64 [43-87]	63 [42-85]	64 [42-87]
- Sodium (mmol/L)	139 [136-141]	138 [136-141]	139 [136-141]
ECG and Echocardiography			
- Atrial fibrillation (%)	606 (24%)	401 (25%)	205 (23%)
- LV ejection fraction (%)	50 [34-60]	50 [34-60]	50 [33-60]
- LV enddiastolic diameter (mm)	50 [44-57]	50 [45-57]	50 [44-57]
Outpatient Medication			
- ACE-I (%)	1263 (51%)	791 (51%)	472 (54%)
- ARB (%)	502 (20%)	306 (19%)	196 (22%)
- Beta-Blockers (%)	1150 (46%)	709 (42%)	441 (50%)
- Spironolactone (%)	227 (9%)	154 (10%)	73 (8%)
Final Adjudicated Diagnosis			
- Acute Heart Failure (%)	1456 (59%)	929 (59%)	527 (58%)

Table 1B: Baseline Characteristics of Patients with non-cardiac Dyspnea

	All patients (n=1029)	**Day-time (n=648)**	**Evening/Night-time (n=381)**
Demographics			
- Age (yrs)	67 [53-78]	69 [56-78]	65 [53-76]
- Male (%)	555 (53%)	332 (51%)	223 (59%)
Medical history			
- Hypertension (%)	541 (53%)	343 (53%)	198 (52%)
- Diabetes (%)	159 (15%)	91 (14%)	68 (18%)
- Coronary Artery Disease (%)	178 (17%)	110 (17%)	68 (18%)
- CKD (%)	129 (13%)	84 (13%)	45 (12%)
- COPD (%)	421 (41%)	262 (40%)	159 (42%)
Physical exam at ED			
- sBP (mmHg)	139±24	137±24	142±23
- HR (beats/min)	93±20	92±19	93±21
- Temperature (°C)	37.4±0.9	37.4±0.9	37.4±1.0
- Oxygen saturation (%)	94±6	94±6	94±5
- Body weight (kg)	73 [61-86]	72 [60-85]	76 [64-89]
- Body Mass Index (kg/m^2)	26 [22-30]	26 [22-30]	26 [22-30]
- Peripheral edema (%)	220 (22%)	144 (23%)	76 (20%)
- Rales (%)	293 (29%)	195 (31%)	98 (26%)
Laboratory parameters			
- Hemoglobin (g/L)	138 [126-150]	137 [124-149]	139 [130-151]
- White Blood Count (g/L)	9.4 [7.1-12.6]	9.6 [7.2-12.6]	9.1 [7.0-12.6]
- C-reactive protein (mg/L)	12.3 [3.0-60.9]	15.5 [3.6-66.3]	9.0 [2.9-51.0]
- Creatinine (mmol/L)	75 [61-93]	73 [59-93]	77 [65-194]
- eGFR (ml/min)	84 [63-101]	84 [63-102]	84 [64-101]
- Sodium (mmol/L)	138 [135-140]	138 [135-140]	138 [136-140]
ECG and Echocardiography			
- Atrial fibrillation (%)	56 (5%)	41 (6%)	15 (4%)
- LV ejection fraction (%)	60 [55-64]	60 [55-63]	60 [55-64]
- LV enddiastolic diameter (mm)	45 [41-50]	45 [40-50]	45 [41-49]
Outpatient Medication			
- ACE-I (%)	219 (21%)	141 (22%)	78 (20%)
- ARB (%)	144 (14%)	88 (14%)	56 (15%)
- Beta-Blockers (%)	252 (24%)	153 (24%)	99 (26%)
- Spironolactone (%)	41 (4%)	33 (5%)	8 (2%)

Diurnal Variations of NP Levels

Among patients adjudicated to have AHF, BNP, NT-proBNP and MR-proANP concentrations were comparable among day-time versus evening/night presenters (all p=ns, Figure 1). In contrast, among patients adjudicated to have other causes of dyspnoea, evening/night-presenters had lower BNP (median 44 ng/L [18-110] versus 75 ng/L [27-168], p<0.01) and NT-proBNP (median 167 ng/L [58-533] versus 269 ng/L [92-827], p<0.01) concentrations versus day-time presenters (Figure 2). This observation was independent of body mass (BNP lean: median 50 ng/L [18-108] versus 70 ng/L [31-181]; BNP obese: 40 nl/L [17-102] versus 79 ng/L [20-226]; NT-proBNP lean: median 156 ng/L [57-522] versus 333 ng/L [114-933] versus NT-proBNP obese: 162 ng/L [60-553] versus 244 ng/L [60-646]. This difference was not observed for MR-proANP concentrations (p=0.52).

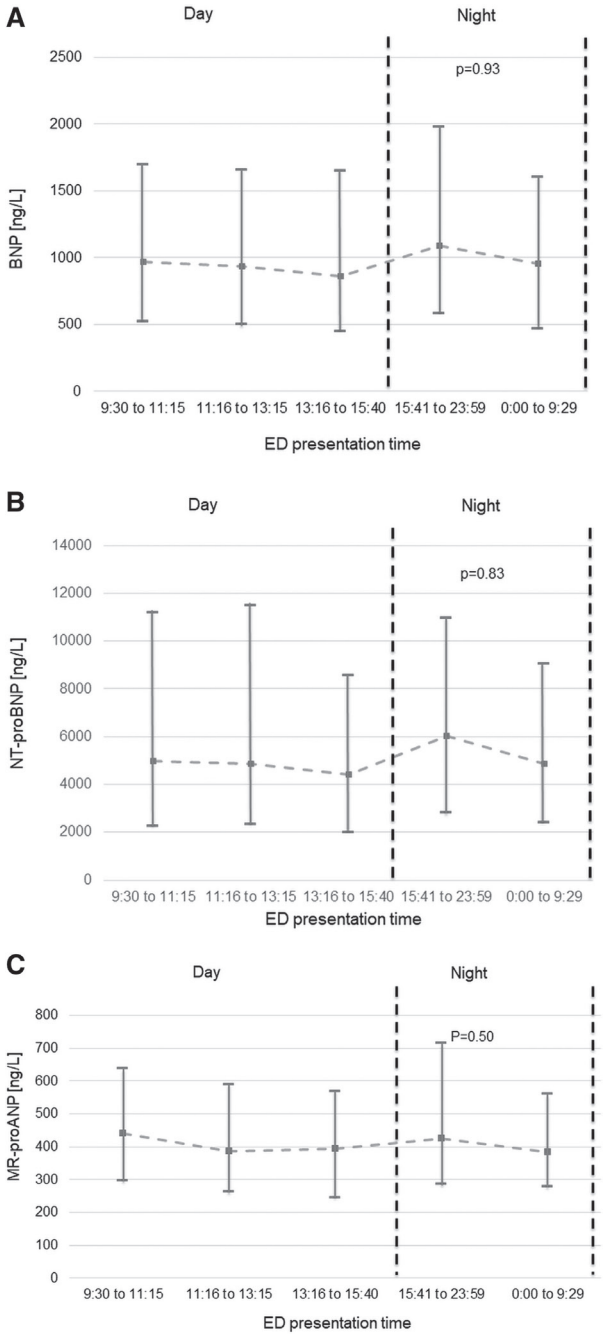


Figure 1: Whisker Plots displaying BNP (A), NT-proBNP (B) and MR-proANP (C) concentrations in AHF patients according to quintiles of presentation time. P values are derived from Jonckheere-Terpstra trend test. Results are displayed as median concentrations, Whisker bars display interquartile ranges.

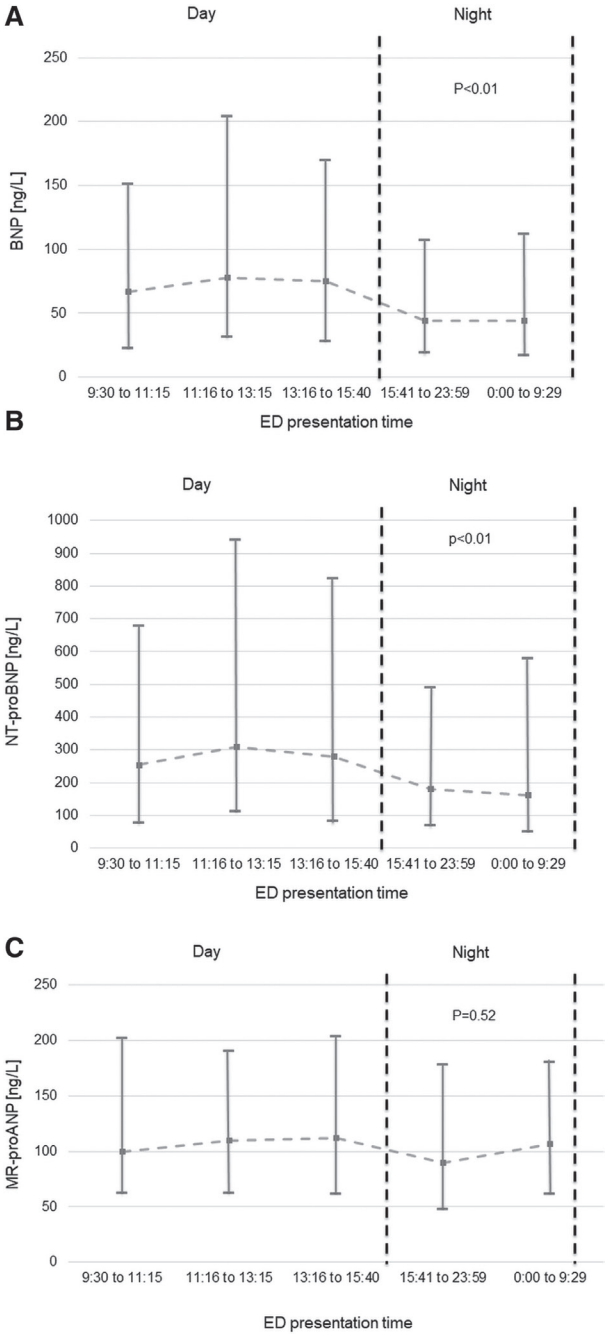


Figure 2: Whisker Plots displaying BNP (A), NT-proBNP (B) and MR-proANP (C) concentrations in patients with non-cardiac dyspnoea according to quintiles of presentation time. P-values are derived from Jonckheere-Terpstra trend test. Results are displayed as median concentrations, Whisker bars display interquartile ranges.

Diagnostic Performance of NPs in Day-time and Evening/Night-time Presenters

Diagnostic accuracy for AHF as quantified by the AUC was significantly higher for BNP (Figure 3A) and NT-proBNP (Figure 3B) in evening/night-time presenters compared to day-time presenters (0.97 [95%CI 0.95-0.98] and 0.95 [95%CI 0.93-0.96]) versus (0.94 [95%CI 0.92-0.95] and 0.91 [95%CI 0.90-0.93]). Secondary analyses including patients with a complete pair of BNP and NT-proBNP samples revealed similar results. Likewise, the diagnostic accuracy for AHF was higher for BNP (lean AUC evening/night: 0.98 versus day-time:0.95; obese AUC evening/night:0.93 versus day-time: 0.89) and NT-proBNP (lean AUC evening/night: 0.96 versus day-time: 0.92; obese AUC evening/night: 0.91 versus day-time: 0.88) in evening/night-time presenters compared to day-time presenters independent of body mass index. The AUC for MR-proANP to detect AHF as the main cause of dyspnoea was similar in evening/night-time and day-time presenters (AUC: 0.92 [95%CI 0.90-0.94] vs. 0.90 [95%CI 0.89-0.92]). Importantly, the diagnostic accuracy of BNP, NT-proBNP and MR-proANP for AHF remained very high (>0.9) at all times. No differences between evening/night-time versus day-time presenters existed for the sensitivity of the three NPs at the guideline recommended rule out cut-off levels at 100ng/L (98% vs. 97%), 300ng/L (98% vs. 98%) and 120pmol/L (97% vs. 97%) respectively. Table 2 displays diagnostic test characteristics of BNP and NT-proBNP at pre-specified sensitivity and specificity target levels in the overall cohort as well as evening/night-time versus day-time presenters.

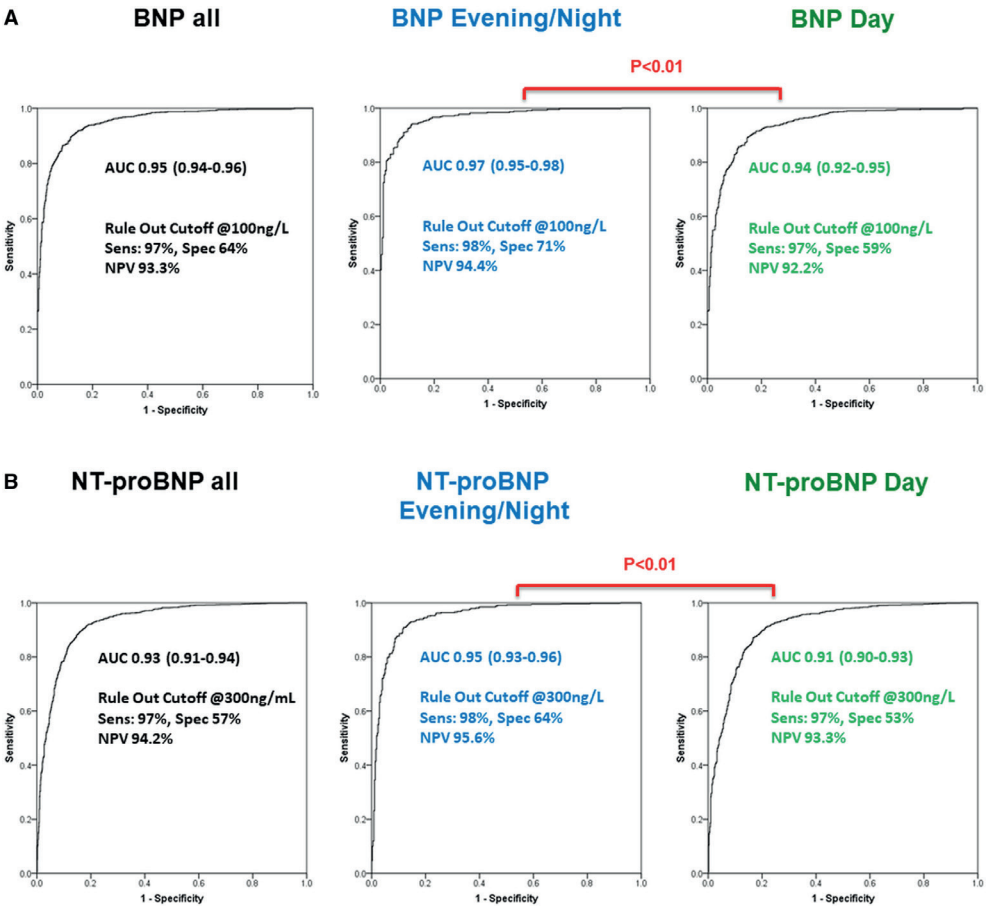


Figure 3: ROC curves displaying the diagnostic accuracy of BNP (A) and NT-proBNP (B) for AHF in day-time and night-time presenters. P-values are derived from comparison of areas under independent ROC curves as recommended by Hanley et al. [20].

Table 2A. Diagnostic test characteristics of BNP at pre-specified sensitivity and specificity targets

	Cut point for Overall	Sensitivity (95% CI)	Specificity (95% CI)	Cut point for Day-time	Sensitivity (95% CI)	Specificity (95% CI)	Cut point for Evening/ Night-time	Sensitivity (95% CI)	Specificity (95% CI)
Target sensitivity									
70%	588ng/L	70% (67.4-72.6)	97% (95.0-97.8)	578ng/L	70% (66.7-73.2)	95% (92.5-96.8)	630ng/L	70% (65.5-74.2)	99% (96.7-99.8)
80%	423ng/L	80% (77.7-82.3)	94% (91.5-95.3)	410ng/L	80% (77.0-82.8)	91% (87.5-93.1)	430ng/L	80% (76.1-83.7)	98% (95.1-99.2)
90%	252ng/L	90% (88.3-91.7)	87% (84.1-89.2)	254ng/L	90% (87.7-92.0)	84% (80.0-87.0)	245ng/L	90% (87.0-92.8)	92% (87.6-94.7)
95%	148ng/L	95% (93.7-96.2)	76% (72.7-79.1)	142ng/L	95% (93.2-96.4)	70% (65.4-74.1)	159ng/L	95% (92.6-96.9)	84% (79.0-88.2)
Target specificity									
70%	127ng/L	96% (95.1-97.3)	70% (66.5-73.3)	145ng/L	95% (93.1-96.3)	70% (65.7-74.3)	95ng/L	98% (95.9-98.9)	70% (64.3-75.7)
80%	179ng/L	94% (92.0-94.9)	80% (76.9-82.9)	219ng/L	92% (89.7-93.6)	80% (76.0-83.6)	132ng/L	97% (94.5-98.1)	80% (74.8-84.8)
90%	326ng/L	86% (83.7-87.7)	90% (87.7-92.2)	397ng/L	82% (78.7-84.2)	90% (87.0-92.7)	219ng/L	91% (88.3-93.7)	90% (85.8-93.4)
95%	501ng/L	76% (73.2-78.1)	95% (93.1-96.5)	579ng/L	70% (66.6-73.1)	95% (92.5-96.8)	328ng/L	85% (81.8-88.6)	95% (91.7-97.3)

Table 2B. Diagnostic test characteristics of NT-proBNP at pre-specified sensitivity and specificity targets

	Cut point for Overall	Sensitivity (95% CI)	Specificity (95% CI)	Cut point for Day-time	Sensitivity (95% CI)	Specificity (95% CI)	Cut point for Evening/ Night-time	Sensitivity (95% CI)	Specificity (95% CI)
Target sensitivity									
70%	2658ng/L	70% (67.7-72.5)	96% (90.7-94.1)	2512ng/L	70% (69.8-72.9)	91% (88.1-92.8)	3046ng/L	70% (65.8-74.0)	96% (93.1-97.5)
80%	1816ng/L	80% (77.9-82.2)	89% (86.9-90.9)	1688ng/L	80% (77.3-82.6)	87% (84.1-89.5)	2120ng/L	80% (76.1-83.3)	93% (90.2-95.6)
90%	1023ng/L	90% (88.4-91.6)	82% (79.5-84.4)	1009ng/L	90% (87.9-91.9)	79% (75.4-81.9)	1146ng/L	90% (87.1-92.5)	88% (84.7-91.5)
95%	567ng/L	95% (93.8-96.2)	71% (67.8-73.5)	536ng/L	95% (93.4-96.4)	67% (63.1-70.5)	610ng/L	95% (92.8-96.8)	77% (72.3-81.1)
Target specificity									
70%	536ng/L	95% (94.1-96.5)	70% (67.2-72.9)	627ng/L	94% (92.4-95.6)	70% (66.1-73.4)	417ng/L	96% (94.2-97.7)	70% (65.0-74.5)
80%	941ng/L	91% (89.8-92.8)	80% (77.5-82.5)	1070ng/L	89% (86.7-90.9)	80% (76.7-83.1)	752ng/L	94% (91.6-95.9)	80% (75.7-84.0)
90%	2016ng/L	77% (75.2-79.6)	90% (88.0-91.8)	2320ng/L	72% (69.0-75.0)	90% (87.4-92.2)	1360ng/L	87% (84.0-90.1)	90% (86.6-92.9)
95%	3766ng/L	60% (57.0-62.3)	95% (93.5-96.3)	4597ng/L	51% (47.6-54.3)	95% (93.1-96.6)	2783ng/L	72% (67.6-75.7)	95% (92.1-96.9)

External validation in Maastricht Serial Sampling Cohort

In most individuals (35 out of 40), we found rhythmic diurnal BNP variation, characterized by the highest concentrations through day-time (mean 21.3 ng/L at 15.30 PM and mean 31.4 ng/L at 14.30 PM for non-CKD and CKD subjects, respectively) and lowest concentrations during night-time (mean 16 ng/L at 4:30 AM and mean 21.4 ng/L at 5:30 AM the next day, respectively—Figure 4A; individual curves in Supplemental Figures 2 and 3). These observed patterns correlated significantly to the fitted cosine model (range R^2 0.18 - 0.83; all P <0.05—see Supplemental Tables 1 and 2). Additionally, we found similar rhythmic diurnal NT-proBNP variation in most individuals (38 out of 40), again characterized by the highest concentrations throughout day-time (mean 292.8 ng/L at 13:30 PM and mean 649.4 ng/L at 13:30 PM for non-CKD and CKD subjects, respectively) and lower concentrations during morning-time (mean 227.2 ng/L at 9:30 AM and mean 504.0 ng/L at 9:30 AM, respectively —Figure 4B; individual curves in Supplemental Figures 4 and 5). These observed patterns correlated significantly to the fitted cosine model (range R^2 0.27 – 0.99; all P ≤0.05—see Supplemental Tables 3 and 4). No significant differences for BNP versus NT-proBNP were found between subjects with or without CKD and between lean and obese patients (Figure 5). There was substantial overlap in phase shifts between lean and obese patients (BNP: 1.75 [95%CI 1.55.-1.95] vs.1.59 [95%CI 1.41-1.77]; NT-proBNP: 1.30 [95%CI 1.16-1.44] vs. 1.04 [95%CI 0.84-1.24]; corresponding to a mean phase difference of 6 minutes for BNP and 16 minutes for NT-proBNP between lean and obese patients. For all of the diurnal rhythms, correction for possible posture-induced changes in plasma volume did not abrogate the diurnal rhythm (data not shown).

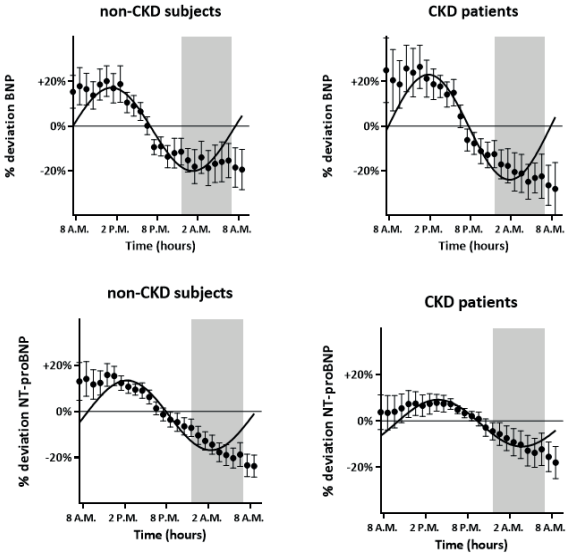


Figure 4: Diurnal variation of BNP (A) and NT-proBNP (B) in non-dyspneic participants according to renal function.

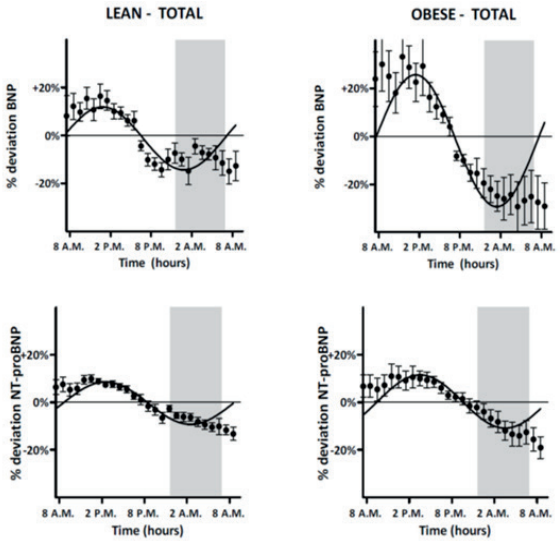


Figure 5: Diurnal variation of BNP (A) and NT-proBNP (B) in non-dyspneic participants according to body mass classes.

Discussion

In a large, multicenter diagnostic study using central adjudication we quantified diurnal variations in NP concentrations and their possible effect on the diagnostic accuracy for AHF among evening/night time versus day-time presenters. We report seven major findings.

First, BNP and NT-proBNP, the two most widely used NPs, exhibit a diurnal rhythm in patients adjudicated to have non-cardiac causes of acute dyspnoea, which is characterized by peak levels around midday and trough levels at night. This confirms previous pilot studies in healthy individuals.[9, 11] Second, the extent of the diurnal variation was moderate with peak/trough differences of 34 ng/L and 114 ng/L for BNP and NT-proBNP, respectively. Third, no clinically relevant impact of body mass on the diurnal rhythm of NP concentrations was observed. Fourth, in AHF patients no diurnal rhythm was observed, possibly because AHF-induced hemodynamic stress overshadowed the moderate naturally occurring day-time/night-time variations. Fifths, the diurnal rhythm of circulating BNP and NT-proBNP concentrations in patients with non-cardiac dyspnoea significantly affected the diagnostic accuracy of these NPs for AHF, with an even higher AUCs for BNP and NT-proBNP in evening/nigh-time presenters (due to lower concentrations in non-AHF patients) versus day-time presenters. Importantly, the diagnostic accuracy of BNP and NT-proBNP for AHF remained very high (AUC >0.9) at all times and the accurate rule-out of AHF at the clinically used cut-off levels of 100 ng/L and 300 ng/L was not affected by presentation time (sensitivity ≥ 97%). Sixth, external validation using hourly measurements of BNP and NT-proBNP in volunteers confirmed a diurnal rhythm for BNP and NT-proBNP with substantially lower concentrations during the night. Reduction in stable body-posture induced plasma volume changes did not underlie these diurnal variation of circulating BNP and NT-proBNP concentrations, suggesting diurnal variations in cardiac filling pressure independent NP release mechanisms. Moreover, no significant differences were found between subjects with or without CKD. An in-vivo study investigating patients awakening from deep anesthesia during off-pump coronary artery bypass grafting found a substantial 300% increase in plasma BNP levels over the first 18 hours following surgery despite constant pulmonary capillary wedge pressures [24], suggesting sympathetic activity as a possible modulator or even inductor of BNP secretion [25]. In fact, a stimulatory effect of adrenergic stimulation on NP secretion is well described [26] and sympathetic nervous system activity itself follows a distinct diurnal pattern, with minimal activity in the second half of the night [27]. Hence, changes in sympathetic nervous system activity might well underlie the diurnal variations observed for BNP and NT-proBNP concentrations in the current study. Seventh, MR-proANP concentrations did not display statistically significant diurnal variations in patients with non-cardiac dyspnoea and the diagnostic accuracy of MR-proANP for AHF was similar in day-time and evening/night-time presenters. Interestingly, a series of in vitro studies using rat myocytes suggest the time course of ANP and BNP

release following sympathetic activation to differ considerably. mRNA concentrations of BNP appeared to rise sharply within the first hour and peak 4 hours after stimulation, while mRNA concentrations of ANP increased only 6-8 hours after stimulation and continued to increase for up to 24 hours [26, 28, 29]. It is therefore perceivable that the delayed ANP response to sympathetic activation might blunt a possible effect of sympathetic activation on its diurnal rhythm. In parallel, differences in the half-life of BNP (20 min) versus NT-proBNP (120 min)[30] might explain the relatively smaller amplitude of the diurnal NT-proBNP variations versus BNP variation observed in this study. Our findings are supported by experimental studies documenting that ANP mRNA did not follow a diurnal oscillation in the mouse heart, whereas BNP mRNA followed a pronounced diurnal rhythm with lowest levels during the 12 hour dark period [31], similar to the variations observed in this study.

Study strengths and limitations

This study has important methodological strengths including its large sample size, its highly representative population of patients presenting to the ED with acute dyspnoea [1, 2], central adjudicated final diagnosis and external validation in a well-described cohort of stable participants.

This study also has a number of limitations. First, MR-proANP concentrations were only available in a subset of patients in the BASEL V ED cohort and not in the Maastricht stable participant cohort. However, sensitivity analyses including only patients with a full set of BNP/NT-proBNP and MR-proANP samples confirmed the results of the overall cohort. However, we cannot comment on any potential diurnal variations in MR-proANP concentrations in healthy participants following a normal wake-sleep pattern. Second, although using a very strong methodology for the central adjudication of the final diagnosis, a very small number of patients may still have been misclassified as either AHF or non-AHF in this study. It is extremely unlikely that this inherent limitation would have influenced the main findings. Third, while enrollment was independent from renal function and a substantial number of patients with renal dysfunction were included in both cohorts, this study did not include patients with terminal kidney failure on chronic hemodialysis. Accordingly, we cannot comment on diurnal variations in NP concentrations in this vulnerable patient population.

In conclusion, in healthy volunteers and in ED patients with non-cardiac causes of acute dyspnoea, BNP and NT-proBNP concentrations exhibit significant diurnal variations, characterized by peak levels around midday and trough levels at night. Despite statistically impacting the diagnostic accuracy for AHF depending on presentation time, the diagnostic accuracy of BNP and NT-proBNP for AHF remains very high (>0.9) at all times. Importantly, the accurate rule-out of AHF at the clinically used cut-off levels of 100 ng/L and 300 ng/L is independent of presentation time (sensitivity ≥ 97% at all times). MR-proANP concentrations do not exhibit diurnal variations.

References

1. Mebazaa A, Yilmaz MB, Levy P, et al. Recommendations on pre-hospital and early hospital management of acute heart failure: a consensus paper from the Heart Failure Association of the European Society of Cardiology, the European Society of Emergency Medicine and the Society of Academic Emergency Medicine--short version. Eur Heart J 2015; 36: 1958-66.
2. Ponikowski P, Voors AA, Anker SD, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2016; 18: 891-975.
3. Jessup M, Drazner MH, Book W, et al. 2017 ACC/AHA/HFSA/ISHLT/ACP Advanced Training Statement on Advanced Heart Failure and Transplant Cardiology (Revision of the ACCF/AHA/ACP/HFSA/ISHLT 2010 Clinical Competence Statement onManagement of Patients With Advanced Heart Failure and CardiacTransplant): A Report of the ACC Competency Management Committee. Journal of the American College of Cardiology 2017; 69: 2977-3001.
4. Januzzi JL, Jr., Camargo CA, Anwaruddin S, et al. The N-terminal Pro-BNP investigation of dyspnea in the emergency department (PRIDE) study. Am J Cardiol 2005; 95: 948-54.
5. Mueller C, Scholer A, Laule-Kilian K, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 2004; 350: 647-54.
6. Maisel A, Mueller C, Nowak R, et al. Mid-region pro-hormone markers for diagnosis and prognosis in acute dyspnea: results from the BACH (Biomarkers in Acute Heart Failure) trial. J Am Coll Cardiol 2010; 55: 2062-76.
7. Mueller C, McDonald K, de Boer RA, et al. Heart Failure Association of the European Society of Cardiology practical guidance on the use of natriuretic peptide concentrations. Eur J Heart Fail 2019; 21: 715-31.
8. Bruins S, Fokkema MR, Romer JW, Dejongste MJ, van der Dijs FP, van den Ouweland JM, Muskiet FA. High intraindividual variation of B-type natriuretic peptide (BNP) and amino-terminal proBNP in patients with stable chronic heart failure. Clin Chem 2004; 50: 2052-8.
9. Goetze JP, Jorgensen HL, Sennels HP, Fahrenkrug J. Diurnal plasma concentrations of natriuretic propeptides in healthy young males. Clin Chem 2012; 58: 789-92.
10. Sothern RB, Vesely DL, Kanabrocki EL, et al. Blood pressure and atrial natriuretic peptides correlate throughout the day. Am Heart J 1995; 129: 907-16.
11. Parcha V, Patel N, Gutierrez OM, et al. Chronobiology of Natriuretic Peptides and Blood Pressure in Lean and Obese Individuals. J Am Coll Cardiol 2021; 77: 2291-303.
12. Wussler D, Kozhuharov N, Sabti Z, et al. External Validation of the MEESSI Acute Heart Failure Risk Score: A Cohort Study. Annals of internal medicine 2019; 170: 248-56.
13. Wussler D, Kozhuharov N, Tavares Oliveira M, et al. Clinical Utility of Procalcitonin in the Diagnosis of Pneumonia. Clinical chemistry 2019; 65: 1532-42.
14. Kozhuharov N, Sabti Z, Wussler D, et al. Prospective validation of N-terminal pro B-type natriuretic peptide cut-off concentrations for the diagnosis of acute heart failure. European journal of heart failure 2019; 21: 813-5.
15. Breidthardt T, Moreno-Weidmann Z, Uthoff H, et al. How accurate is clinical assessment of neck veins in the estimation of central venous pressure in acute heart failure? Insights from a prospective study. European journal of heart failure 2018; 20: 1160-2.
16. Klinkenberg LJ, van Dijk JW, Tan FE, van Loon LJ, van Dieijen-Visser MP, Meex SJ. Circulating cardiac troponin T exhibits a diurnal rhythm. J Am Coll Cardiol 2014; 63: 1788-95.
17. Klinkenberg LJ, Wildi K, van der Linden N, et al. Diurnal Rhythm of Cardiac Troponin: Consequences for the Diagnosis of Acute Myocardial Infarction. Clin Chem 2016; 62: 1602-11.
18. van der Linden N, Cornelis T, Kimenai DM, et al. Origin of Cardiac Troponin T Elevations in Chronic Kidney Disease. Circulation 2017; 136: 1073-5.
19. Inker LA, Schmid CH, Tighiouart H, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. N Engl J Med 2012; 367: 20-9.

20. Morgenthaler NG, Struck J, Thomas B, Bergmann A. Immunoluminometric assay for the midregion of pro-atrial natriuretic peptide in human plasma. Clin Chem 2004; 50: 234-6.
21. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983; 148: 839-43.
22. Nelson W, Tong YL, Lee JK, Halberg F. Methods for cosinor-rhythmometry. Chronobiologia 1979; 6: 305-23.
23. Tong YL. Parameter estimation in studying circadian rhythms. Biometrics 1976; 32: 85-94.
24. Puelacher C, Rudez J, Twerenbold R, et al. B-type natriuretic peptide secretion without change in intra-cardiac pressure. Clinical biochemistry 2015; 48: 318-21.
25. Valette X, Lemoine S, Allouche S, Gerard JL, Hanouz JL. Effect of lipopolysaccharide, cytokines, and catecholamines on brain natriuretic peptide release from human myocardium. Acta Anaesthesiol Scand 2012; 56: 860-5.
26. Luchner A, Schunkert H. Interactions between the sympathetic nervous system and the cardiac natriuretic peptide system. Cardiovasc Res 2004; 63: 443-9.
27. Burgess HJ, Trinder J, Kim Y. Cardiac autonomic nervous system activity during presleep wakefulness and stage 2 NREM sleep. J Sleep Res 1999; 8: 113-22.
28. Hanford DS, Glembotski CC. Stabilization of the B-type natriuretic peptide mRNA in cardiac myocytes by alpha-adrenergic receptor activation: potential roles for protein kinase C and mitogen-activated protein kinase. Mol Endocrinol 1996; 10: 1719-27.
29. Hanford DS, Thuerauf DJ, Murray SF, Glembotski CC. Brain natriuretic peptide is induced by alpha 1-adrenergic agonists as a primary response gene in cultured rat cardiac myocytes. J Biol Chem 1994; 269: 26227-33.
30. de Lemos JA, McGuire DK, Drazner MH. B-type natriuretic peptide in cardiovascular disease. Lancet (London, England) 2003; 362: 316-22.
31. Goetze JP, Georg B, Jorgensen HL, Fahrenkrug J. Chamber-dependent circadian expression of cardiac natriuretic peptides. Regul Pept 2010; 160: 140-5.

Supplemental material

Tables

Supplemental Table 1: Baseline characteristics of subjects without CKD (n=19) and subjects with CKD (n=21).

	Subjects without CKD (n=19)	Subjects with CKD (n=21)
Age (years)	72 ± 7	66 ± 12
Male sex	15 (79)	14 (67)
Body mass index (kg/m^2)	26 ± 3	28 ± 4
Diabetes Mellitus	6 (32)	7 (33)
Cholesterol concentration (mg/dL)	176 ± 35	157 ± 33
Systolic blood pressure (mmHg)	140 ± 15	136 ± 19
Diastolic blood pressure (mmHg)	68 ± 8	86 ± 14
Creatinine (mg/dL)	1.0 ± 0.2	3.3 ± 1.0
Cystatin C (mg/L)	1.0 ± 0.2	2.8 ± 0.8
MDRD (ml/min/1.73m^2)	73.4 ± 18.5	19.2 ± 6.4
CKD-EPI (ml/min/1.73m^2)	72.9 ± 17.2	18.9 ± 6.6
Chronic Kidney Disease	0 (0)	100 (21)

Continuous data is presented as mean ± SD and categorical data is presented as n (%).

Supplemental Table 2. The individual diurnal variation of BNP in non CKD-subjects is significantly described by a cosinor model.[a]

Participant[b]	Mesor (ng/L)	Amplitude (ng/L)	Acrophase (h)	R^2	P-Value
1	11,8	3,2	13.15 PM	0,80	<0.001
2	3,0	0,1	10.04 AM	0,40	0.05
3	5,9	0,5	16.51 PM	0,52	<0.001
4	7,9	1,5	15.20 PM	0,74	<0.001
5	3,0	0,1	10.59 AM	0,53	<0.05
6	3,9	0,8	10.52 AM	0,77	<0.001
7	25,0	7,4	15.13 PM	0,48	<0.001
8	8,8	3,4	15.43 PM	0,69	<0.001
9	30,3	5,6	15.18 PM	0,59	<0.001
10	3,9	0,7	15.28 PM	0,67	<0.001
11	19,9	4,3	13.10 PM	0,72	<0.001
12	7,0	2,3	14.34 PM	0,60	<0.001
13	32,4	5,2	14.14 PM	0,64	<0.001
14	7,9	0,1	16.30 PM	0,02	0.05
15	76,9	8,9	13.51 PM	0,60	<0.001
16	4,4	1,4	12.30 PM	0,71	<0.001
17	31,4	3,7	19.11 PM	0,62	0.001
18	6,9	3,4	15.12 PM	0,75	<0.001
19	29,6	2,6	13.57 PM	0,59	<0.001
20	3,6	0,5	01.33 AM	0,73	0.05
21	12,7	3,8	15.31 PM	0,64	<0.001

[a] The cosinor model is described as: $Z(t) = M + A \cdot \cos(\omega t + \varphi) + e(t)$, where Z(t) represents the measured hs-cTnT concentration at a given time (t), M the mesor (value about which oscillation occurs), A the Amplitude (half the difference between the peak and the nadir value), ω the angular frequency (degrees per unit time, with 360° representing a complete cycle), φ the acrophase (timing of maximal value in degrees) and e(t) the error between the cosinor model and the measurement.

[b] The numbering used in this table is extended to Supplemental Figure 1, allowing direct comparisons throughout the paper.

Supplemental Table 3. The individual diurnal variation of BNP in CKD-patients is significantly described by a cosinor model.[a]

Participant[b]	Mesor (ng/L)	Amplitude (ng/L)	Acrophase (h)	R^2	P-Value
1	5,1	0,4	15.31 PM	0,02	0.05
2	6,6	3,1	17.26 PM	0,83	<0.001
3	12,3	3,8	15.50 PM	0,58	<0.001
4	4,3	0,9	20.43 PM	0,77	<0.001
5	3,1	0,8	14.56 PM	0,54	0.001
6	25,7	1,0	22.56 PM	0,18	0.001
7	32,5	4,4	15.17 PM	0,60	<0.001
8	11,6	3,9	01.13 AM	0,19	0.05
9	42,1	15,5	15.42 PM	0,61	<0.001
10	11,1	2,5	14.31 PM	0,47	0.001
11	6,4	1,9	15.21 PM	0,56	<0.001
12	5,0	1,0	14.39 PM	0,44	0.001
13	20,6	5,1	16.37 PM	0,57	<0.001
14	9,7	4,5	15.10 PM	0,73	<0.001
15	26,4	9,6	15.37 PM	0,59	<0.001
16	15,6	6,7	14.41 PM	0,51	0.001
17	42,5	13,4	15.13 PM	0,60	<0.001
18	153,5	8,5	11.56 AM	0,43	0.001
19	32,0	7,6	13.53 PM	0,78	<0.001

[a] *The cosinor model is described as: $Z(t) = M + A \cdot \cos(\omega t + \varphi) + e(t)$, where $Z(t)$ represents the measured hs-cTnT concentration at a given time (t), M the mesor (value about which oscillation occurs), A the Amplitude (half the difference between the peak and the nadir value), ω the angular frequency (degrees per unit time, with 360° representing a complete cycle), φ the acrophase (timing of maximal value in degrees) and e(t) the error between the cosinor model and the measurement.*

[b] *The numbering used in this table is extended to Supplemental Figure 2, allowing direct comparisons throughout the paper.*

Supplemental Table 4. The individual diurnal variation of NT-proBNP in non-CKD subjects is significantly described by a cosinor model.[a]

Participant[b]	Mesor (ng/L)	Amplitude (ng/L)	Acrophase (h)	R^2	P-Value
1	281,9	30,8	19.18 PM	0,80	<0.001
2	78,5	20,2	12.41 PM	0,56	<0.001
3	21,5	4,2	13.56 PM	0,36	0.001
4	103,7	29,5	15.06 PM	0,72	<0.001
5	26,7	9,1	11.46 AM	0,62	0.001
6	353,1	36,0	13.40 PM	0,89	<0.001
7	808,8	101,2	20.13 PM	0,92	<0.001
8	73,8	10,2	17.09 PM	0,49	0.001
9	6,1	1,7	12.13 PM	0,48	0.001
10	136,6	32,1	17.00 PM	0,75	<0.001
11	39,3	7,6	14.47 PM	0,40	<0.001
12	334,4	36,5	15.25 PM	0,65	<0.001
13	77,5	5,4	15.43 PM	0,27	<0.001
14	50,1	8,6	11.10 AM	0,99	<0.001
15	125,5	36,7	15.30 PM	0,69	<0.001
16	195,4	44,5	14.22 PM	0,66	<0.001
17	61,6	11,3	12.54 PM	0,69	<0.001
18	32,5	3,6	16.14 PM	0,50	<0.001
19	526,5	105,8	15.58 PM	0,50	<0.001
20	2009,0	255,7	23.33 PM	0,56	0.001
21	28,4	7,3	18.43 PM	0,70	<0.001

[a] The cosinor model is described as: $Z(t) = M + A \cdot \cos(\omega t + \varphi) + e(t)$, where $Z(t)$ represents the measured hs-cTnT concentration at a given time (t), M the mesor (value about which oscillation occurs), A the Amplitude (half the difference between the peak and the nadir value), ω the angular frequency (degrees per unit time, with 360° representing a complete cycle), φ the acrophase (timing of maximal value in degrees) and e(t) the error between the cosinor model and the measurement.

[b] The numbering used in this table is extended to Supplemental Figure 3, allowing direct comparisons throughout the paper

Supplemental Table 5. The individual diurnal variation of NT-proBNP in CKD-patients is significantly described by a cosinor model.[a]

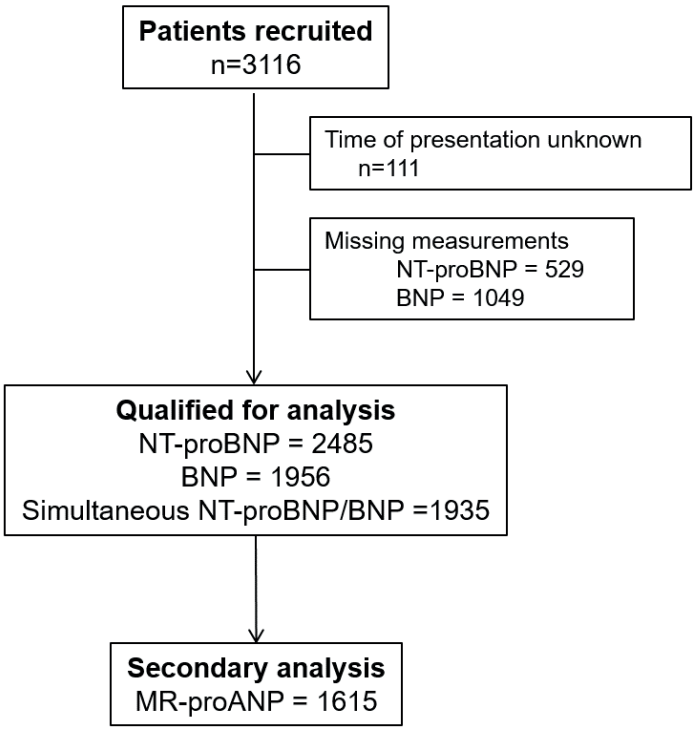
Participant[b]	Mesor (ng/L)	Amplitude (ngl/L)	Acrophase (h)	R^2	P-Value
1	223,5	7,7	18.43 PM	0,41	0.05
2	76,5	28,8	13.46 PM	0,53	<0.001
3	421,4	50,6	16.07 PM	0,53	<0.001
4	72,6	13,1	00.00 AM	0,78	<0.001
5	50,2	7,8	18.04 PM	0,79	<0.001
6	791,5	25,1	21.55 PM	0,62	<0.001
7	1584,0	28,0	23.02 PM	0,11	0.05
8	78,4	8,6	23.02 PM	0,89	<0.001
9	1342,0	251,4	17.04 PM	0,71	<0.001
10	829,8	76,7	16.25 PM	0,67	<0.001
11	156,3	16,3	17.16 PM	0,70	<0.001
12	156,0	18,5	16.15 PM	0,49	<0.001
13	380,7	45,1	16.18 PM	0,64	<0.001
14	497,0	120,6	16.56 PM	0,80	<0.001
15	562,6	113,3	17.40 PM	0,84	<0.001
16	652,2	146,5	15.36 PM	0,54	<0.001
17	694,6	100,0	16.07 PM	0,60	<0.001
18	1320,0	122,1	01.22 AM	0,76	<0.001
19	1889,0	331,8	15.41 PM	0,72	<0.001

[a] The cosinor model is described as: $Z(t) = M + A \cdot \cos(\omega t + \varphi) + e(t)$, where Z(t) represents the measured hs-cTnT concentration at a given time (t), M the mesor (value about which oscillation occurs), A the Amplitude (half the difference between the peak and the nadir value), ω the angular frequency (degrees per unit time, with 360° representing a complete cycle), φ the acrophase (timing of maximal value in degrees) and e(t) the error between the cosinor model and the measurement.

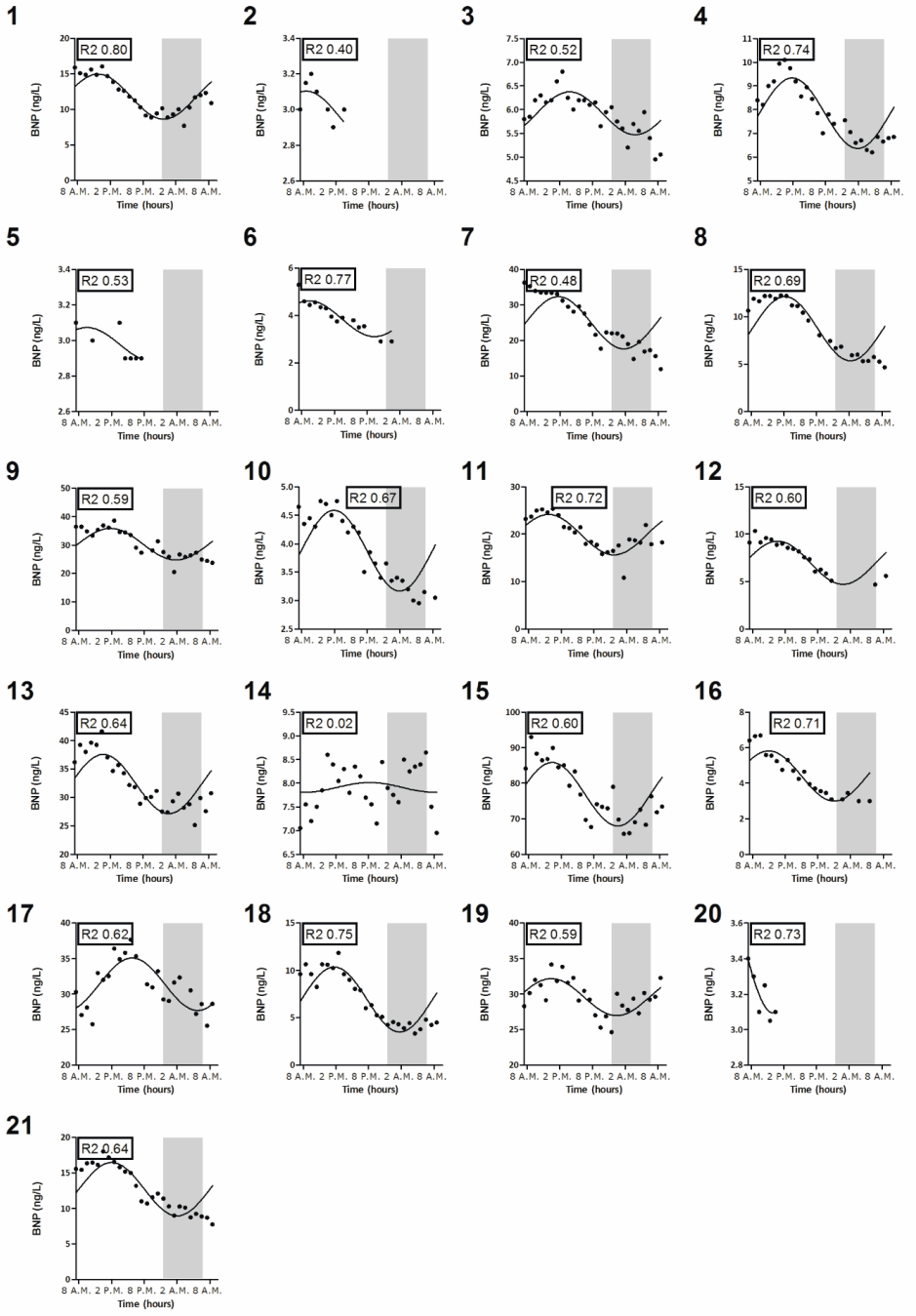
[b] The numbering used in this table is extended to Supplemental Figure 4, allowing direct comparisons throughout the paper

Figures

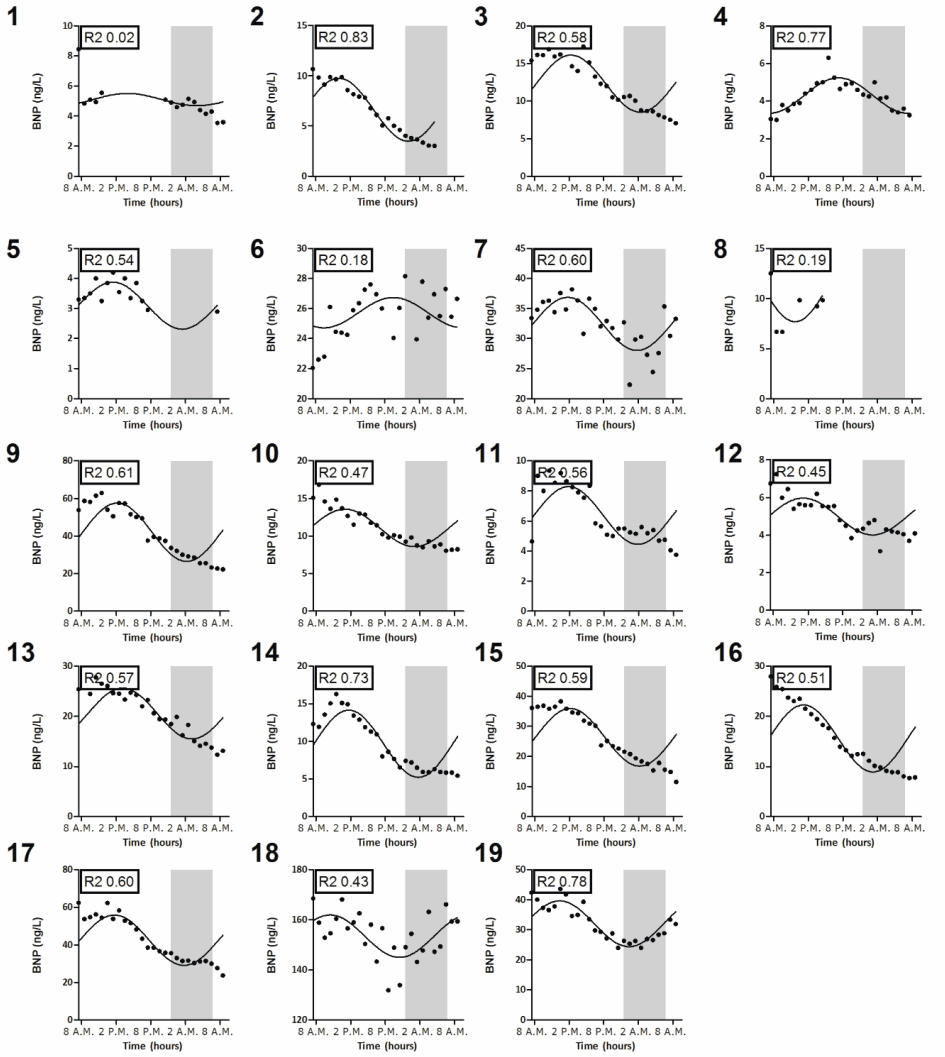
Supplemental Figure 1. Patient flow chart.



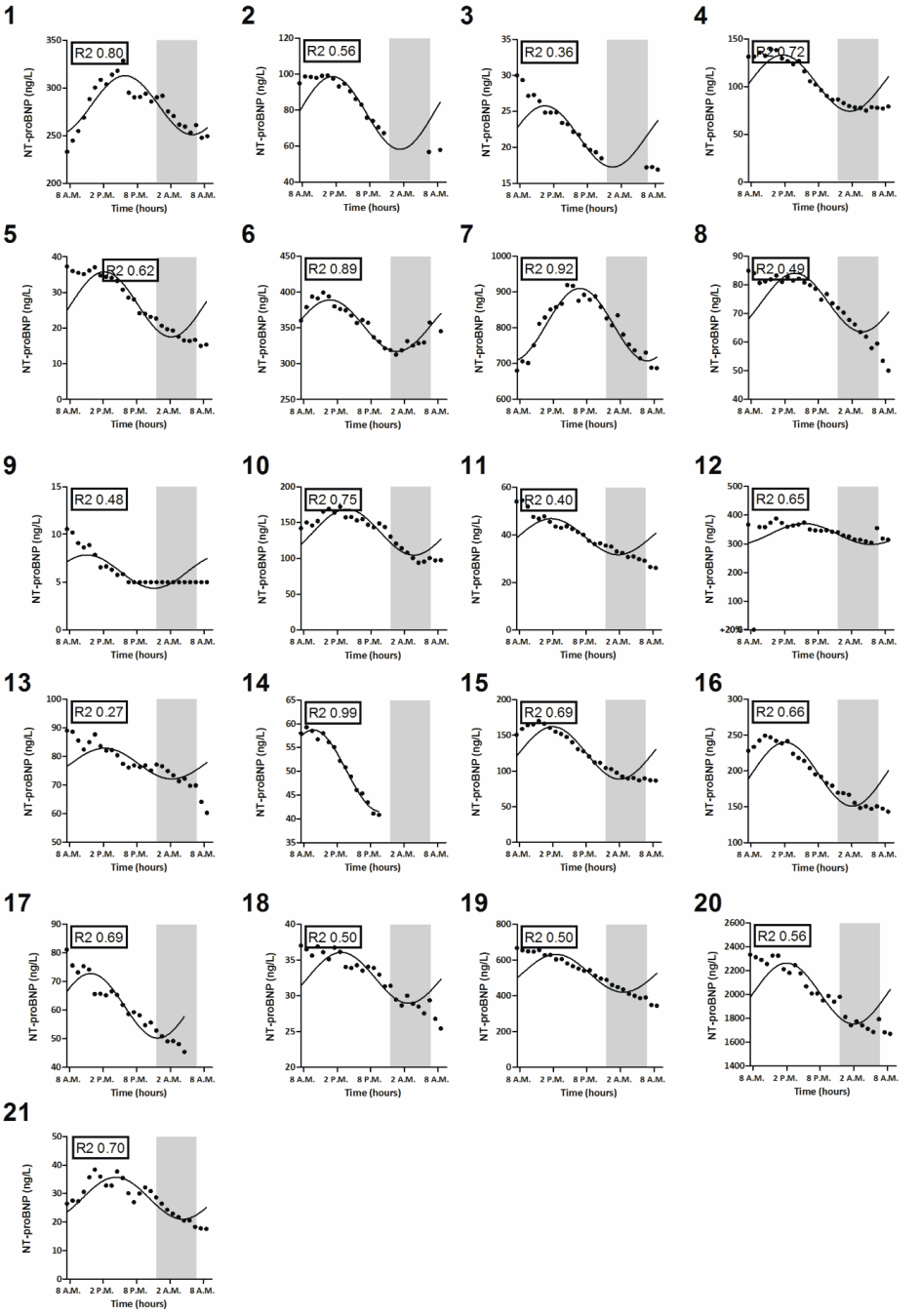
Supplemental figure 2. Individual diurnal BNP profiles of non-CKD subjects and fitted cosine curves.



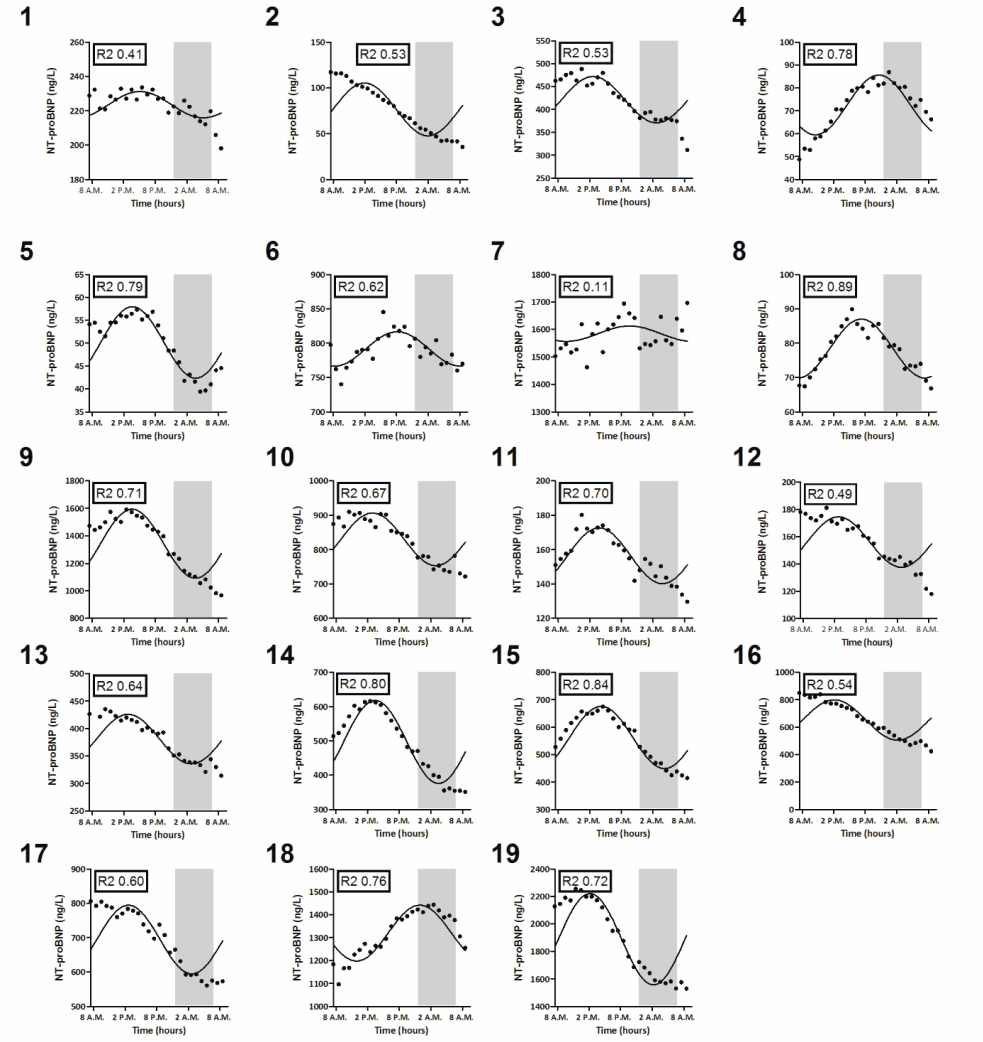
Supplemental figure 3. Individual diurnal BNP profiles of CKD patients and fitted cosine curves.



Supplemental figure 4. Individual diurnal NT-proBNP profiles of non-CKD subjects and fitted cosine curves.



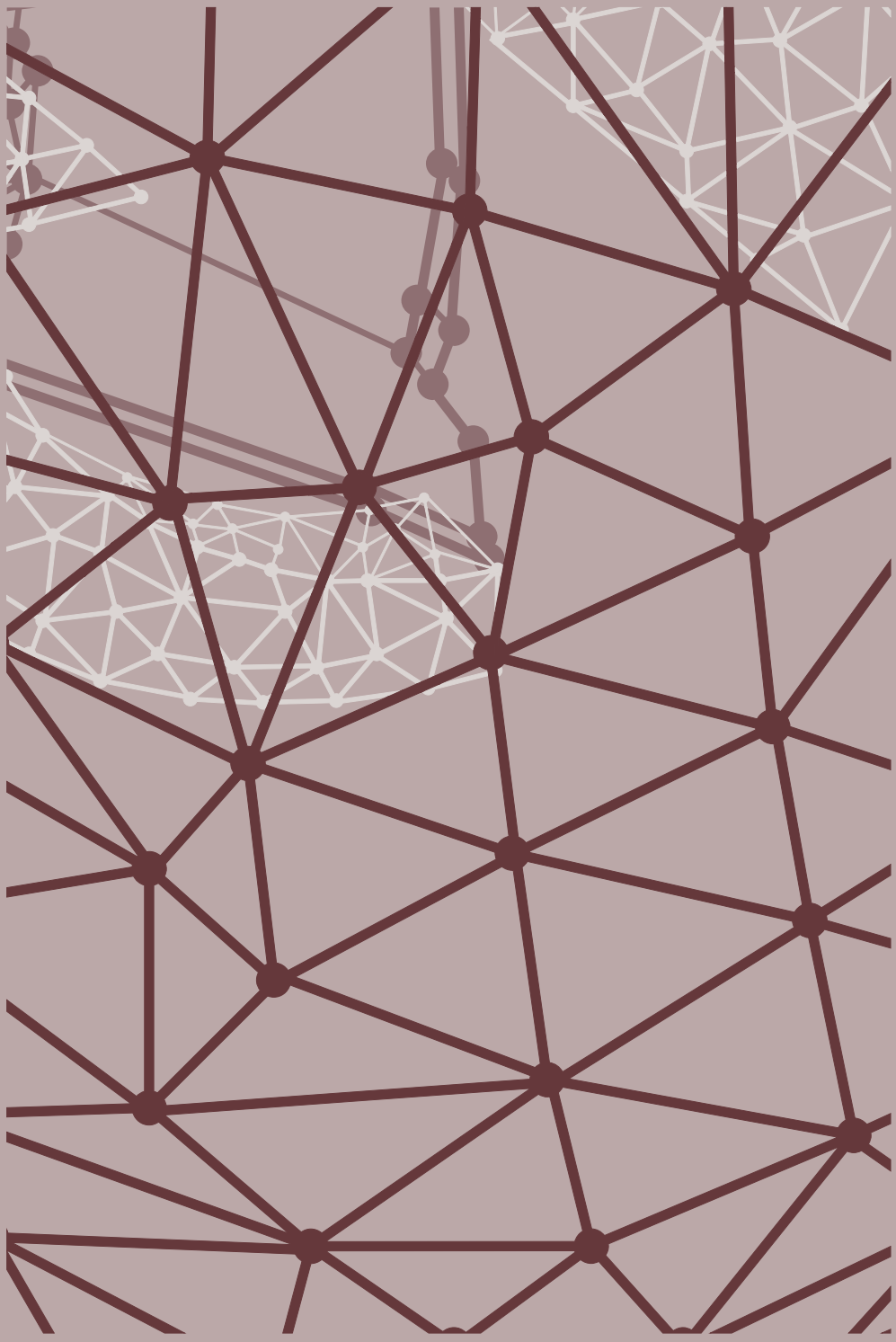
Supplemental figure 5. Individual diurnal NT-proBNP profiles of CKD patients and fitted cosine curves.



Supplemental References

1. Januzzi JL, Jr., Camargo CA, Anwaruddin S, Baggish AL, Chen AA, Krauser DG, Tung R, Cameron R, Nagurney JT, Chae CU, Lloyd-Jones DM, Brown DF, Foran-Melanson S, Sluss PM, Lee-Lewandrowski E and Lewandrowski KB. The N-terminal Pro-BNP investigation of dyspnea in the emergency department (PRIDE) study. Am J Cardiol. 2005;95:948-54.
2. Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, Pfisterer M and Perruchoud AP. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med. 2004;350:647-54.
3. Maisel A, Mueller C, Nowak R, Peacock WF, Landsberg JW, Ponikowski P, Mockel M, Hogan C, Wu AH, Richards M, Clopton P, Filippatos GS, Di Somma S, Anand I, Ng L, Daniels LB, Neath SX, Christenson R, Potocki M, McCord J, Terracciano G, Kremastinos D, Hartmann O, von Haehling S, Bergmann A, Morgenthaler NG and Anker SD. Mid-region pro-hormone markers for diagnosis and prognosis in acute dyspnea: results from the BACH (Biomarkers in Acute Heart Failure) trial. J Am Coll Cardiol. 2010;55:2062-76.
4. Mueller C, McDonald K, de Boer RA, Maisel A, Cleland JGF, Kozhuharov N, Coats AJS, Metra M, Mebazaa A, Ruschitzka F, Lainscak M, Filippatos G, Seferovic PM, Meijers WC, Bayes-Genis A, Mueller T, Richards M, Januzzi JL, Jr. and Heart Failure Association of the European Society of C. Heart Failure Association of the European Society of Cardiology practical guidance on the use of natriuretic peptide concentrations. Eur J Heart Fail. 2019;21:715-731.
5. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, Falk V, Gonzalez-Juanatey JR, Harjola VP, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GM, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P, Authors/Task Force M and Document R. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail. 2016;18:891-975.
6. Jessup M, Drazner MH, Book W, Cleveland JC, Jr., Dauber I, Farkas S, Ginwalla M, Katz JN, Kirkwood P, Kittleson MM, Marine JE, Mather P, Morris AA, Polk DM, Sakr A, Schlendorf KH and Vorovich EE. 2017 ACC/AHA/HFSA/ISHLT/ACP Advanced Training Statement on Advanced Heart Failure and Transplant Cardiology (Revision of the ACCF/AHA/ACP/HFSA/ISHLT 2010 Clinical Competence Statement onManagement of Patients With Advanced Heart Failure and CardiacTransplant): A Report of the ACC Competency Management Committee. Journal of the American College of Cardiology. 2017;69:2977-3001.
7. Bruins S, Fokkema MR, Romer JW, Dejongste MJ, van der Dijs FP, van den Ouweland JM and Muskiet FA. High intraindividual variation of B-type natriuretic peptide (BNP) and amino-terminal proBNP in patients with stable chronic heart failure. Clin Chem. 2004;50:2052-8.
8. Goetze JP, Jorgensen HL, Sennels HP and Fahrenkrug J. Diurnal plasma concentrations of natriuretic propeptides in healthy young males. Clin Chem. 2012;58:789-92.
9. Sothern RB, Vesely DL, Kanabrocki EL, Bremner FW, Third JL, Boles MA, Nemchausky BM, Olwin JH and Scheving LE. Blood pressure and atrial natriuretic peptides correlate throughout the day. Am Heart J. 1995;129:907-16.
10. Parcha V, Patel N, Gutierrez OM, Li P, Gamble KL, Musunuru K, Margulies KB, Cappola TP, Wang TJ, Arora G and Arora P. Chronobiology of Natriuretic Peptides and Blood Pressure in Lean and Obese Individuals. J Am Coll Cardiol. 2021;77:2291-2303.
11. Wussler D, Kozhuharov N, Sabti Z, Walter J, Strebel I, Scholl L, Miro O, Rossello X, Martin-Sanchez FJ, Pocock SJ, Nowak A, Badertscher P, Twerenbold R, Wildi K, Puelacher C, du Fay de Lavallaz J, Shrestha S, Strauch O, Flores D, Nestelberger T, Boeddinghaus J, Schumacher C, Goudev A, Pfister O, Breidthardt T and Mueller C. External Validation of the MEESSI Acute Heart Failure Risk Score: A Cohort Study. Annals of internal medicine. 2019;170:248-256.

12. Wussler D, Kozhuharov N, Tavares Oliveira M, Bossa A, Sabti Z, Nowak A, Murray K, du Fay de Lavallaz J, Badertscher P, Twerenbold R, Shrestha S, Flores D, Nestelberger T, Walter J, Boeddinghaus J, Zimmermann T, Koechlin L, von Eckardstein A, Breidthardt T and Mueller C. Clinical Utility of Procalcitonin in the Diagnosis of Pneumonia. Clinical chemistry. 2019;65:1532-1542.
13. Kozhuharov N, Sabti Z, Wussler D, Nowak A, Badertscher P, Twerenbold R, Wildi K, Stallone F, Vogt F, Hilti J, Puelacher C, du Fay de Lavallaz J, Shrestha S, Flores D, Nestelberger T, Koechlin L, Boeddinghaus J, Zimmermann T, Walter J, Schumacher C, Rentsch K, von Eckardstein A, Keller DI, Goudev A, Pfister O, Breidthardt T, Mueller C, Investigators BV, Osswald S, Reichlin T, Gimenez MR, Sazgary L and Lohrmann J. Prospective validation of N-terminal pro B-type natriuretic peptide cut-off concentrations for the diagnosis of acute heart failure. European journal of heart failure. 2019;21:813-815.
14. Breidthardt T, Moreno-Weidmann Z, Uthoff H, Sabti Z, Aeppli S, Puelacher C, Stallone F, Twerenbold R, Wildi K, Kozhuharov N, Wussler D, Flores D, Shrestha S, Badertscher P, Boeddinghaus J, Nestelberger T, Gimenez MR, Staub D, Aschwanden M, Lohrmann J, Pfister O, Osswald S and Mueller C. How accurate is clinical assessment of neck veins in the estimation of central venous pressure in acute heart failure? Insights from a prospective study. European journal of heart failure. 2018;20:1160-1162.
15. Klinkenberg LJ, van Dijk JW, Tan FE, van Loon LJ, van Dieijen-Visser MP and Meex SJ. Circulating cardiac troponin T exhibits a diurnal rhythm. J Am Coll Cardiol. 2014;63:1788-95.
16. Klinkenberg LJ, Wildi K, van der Linden N, Kouw IW, Niens M, Twerenbold R, Rubini Gimenez M, Puelacher C, Daniel Neuhaus J, Hillinger P, Nestelberger T, Boeddinghaus J, Grimm K, Sabti Z, Bons JA, van Suijlen JD, Tan FE, Ten Kate J, Bekers O, van Loon LJ, van Dieijen-Visser MP, Mueller C and Meex SJ. Diurnal Rhythm of Cardiac Troponin: Consequences for the Diagnosis of Acute Myocardial Infarction. Clin Chem. 2016;62:1602-1611.
17. van der Linden N, Cornelis T, Kimenai DM, Klinkenberg LJJ, Hilderink JM, Luck S, Litjens EJR, Peeters F, Streng AS, Breidthardt T, van Loon LJC, Bekers O, Kooman JP, Westermark PO, Mueller C and Meex SJR. Origin of Cardiac Troponin T Elevations in Chronic Kidney Disease. Circulation. 2017;136:1073-1075.
18. Inker LA, Schmid CH, Tighiouart H, Eckfeldt JH, Feldman HI, Greene T, Kusek JW, Manzi J, Van Lente F, Zhang YL, Coresh J, Levey AS and Investigators C-E. Estimating glomerular filtration rate from serum creatinine and cystatin C. N Engl J Med. 2012;367:20-9.
19. Morgenthaler NG, Struck J, Thomas B and Bergmann A. Immunoluminometric assay for the midregion of pro-atrial natriuretic peptide in human plasma. Clin Chem. 2004;50:234-6.
20. Hanley JA and McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148:839-43.
21. Nelson W, Tong YL, Lee JK and Halberg F. Methods for cosinor-rhythmometry. Chronobiologia. 1979;6:305-23.
22. Tong YL. Parameter estimation in studying circadian rhythms. Biometrics. 1976;32:85-94.



CHAPTER 6

CHARACTERIZATION OF NITROIMIDAZOLE-PROTEIN ADDUCTS: TOWARDS A HYPOXIA-SPECIFIC TROPONIN ASSAY

William P.T.M. van Doorn, Vasantha K. Kadambar, Jella van de Laak, Jordy M.M. Kocken, Alexander S. Streng, Niko Deckers, Leon J. Schurgers, Kasper M.A. Rouschop, Freek G. Bouwman, Artem Melman, Costel C. Darie, Otto Bekers, Cameron J. Koch, Steven J.R. Meex

SUBMITTED FOR PUBLICATION

Abstract

Background: Cardiac troponins are cornerstone in the diagnosis of myocardial infarction (MI). As elevated troponin levels can be associated not only with MI but also with numerous other (non-)cardiac pathologies, increased assay specificity is desirable. In this study we propose the detection of troponin-nitroimidazole adducts as basis for the development of a hypoxia-specific cardiac troponin assay.

Methods: As 2-nitroimidazoles specifically bind to proteins during hypoxia, we used three of them in the current study: pimonidazole, EF5 and DNI. Nitroimidazole binding was initially evaluated in hypoxic HL-1, SV40 and HepG2 cells. Next, albumin-, lysozyme and troponin-nitroimidazole adducts were derived using zinc and radiochemical reduction methods. The exact configuration of the protein-nitroimidazole adduct was analyzed using liquid chromatography coupled with tandem mass-spectrometry (LC-MS/MS).

Results: Hypoxia- and time-dependent accumulation of protein-nitroimidazole adducts was observed in three different cell systems. Protein-nitroimidazole adducts were exclusively formed in availability of the cysteine amino acid. This was confirmed by LC-MS/MS analysis of protein-nitroimidazole adducts, localizing the binding occurring at Cys^{80} for cardiac troponin I.

Conclusions: Our study has confirmed and elucidated the role that cysteine plays in protein-nitroimidazole interactions, providing a basis for development of hypoxia-specific troponin assays.

Introduction

Protein biomarkers are indispensable in medicine for the diagnosis and follow-up of patients. Ideally, measurement of protein biomarkers is both sensitive and specific for damage to the tissue from which they originate. Sometimes, in clinical practice, an additional layer of biomarker specificity would be desirable: the ability to discriminate between hypoxic injury versus other causes of cell death. Examples include patients with chest pain due to acute myocardial infarction versus those with other causes of cardiac injury, the diagnosis of renal ischemia, and the diagnosis of ischemic bowel disease. Rapid identification of hypoxia as the cause of cell death is key for initiating the appropriate treatment and affects prognosis of these patients.

Currently, there are either biomarkers for hypoxia, which lack tissue specificity (e.g. HIF-1-alpha, carbonic anhydrase IX), or markers that reflect tissue injury, but are unable to differentiate between hypoxic injury versus other causes of cell death (e.g. troponin or NGAL)[1-4]. Combining tissue specificity with hypoxia in a single biomarker is an unmet clinical need, but as a native molecule, such marker may not exist.

2-Nitroimidazoles are an interesting family of compounds as they are known to bind proteins during hypoxic stress[5,6]. These exogenous molecules freely diffuse across cell membranes, and -in living cells only- will be metabolized through an enzyme-mediated single electron reduction to form a free radical. This free radical is rapidly reversed to its original compound by intracellular oxygen, which has a higher electron affinity than the original nitro group. However, at decreased intracellular oxygen levels (hypoxia), 2-nitroimidazoles are further reduced resulting in the covalent binding of 2-nitroimidazoles to cellular macromolecules including proteins, DNA and low-level metabolites (e.g., free thiols, glutathione)[7,8]. This feature of 2-nitroimidazoles is employed in nuclear medicine to visualize solid tumors in patients[9]. Unfortunately, the precise configuration of 2-nitroimidazole modifications on proteins remains unproven for more than 40 years. Elucidation of this configuration would allow the site-selective detection of 2-nitroimidazole modifications and provide a technical framework to combine tissue specificity with hypoxia in a single biomarker.

In the present study, we aimed to unravel the precise configuration of the hypoxia-specific 2-nitroimidazole modification on proteins. Initially, we demonstrate hypoxia-specific and time-dependent retention of 2-nitroimidazoles in three cellular model systems of hypoxia. Next, we modified several model proteins with various 2-nitroimidazoles to identify cysteine as the anchor for either the C4 or C5 of the imidazole ring from the hydroxylamine derivative of 2-nitroimidazoles to bind through a sulfide bond. Our results elucidate the binding mechanism of 2-nitroimidazoles to proteins, and provide the molecular basis for the development of protein biomarker assays that discriminate between hypoxic and non-hypoxic cell death.

Methods

Chemicals and reagents (supplemental)

All reagents used in the current study were bought from Sigma-Aldrich and at least reagent-grade, unless stated otherwise. We used three distinct 2-nitroimidazoles: pimonidazole (Hypoxyprobe), EF5 (Hypoxia-imaging) and 1-dansylaminopropyl-2-nitroimidazole (DNI; synthesized in-house) (see supplemental information A and B). UPLC-grade water, acetonitrile (ACN), formic acid, and trifluoroacetic acid (TFA) for sample preparation and mass spectrometry were mass spectrometry-grade and obtained from Biosolve.

Cell culture

The cell lines studied included HL-1 (kindly provided by Dr. W. Claycomb, Louisiana State University, New Orleans, LA, USA) [1], SV40 immortalized human cardiomyocytes (ABMGood) [2,3] and HepG2 [4]. HL-1 cells were maintained in Claycomb medium supplemented with 10% fetal bovine serum (FBS; JRH Biosciences), 0.1 mM norepinephrine, 2 mM L-Glutamine (Invitrogen) and 1% penicillin/streptomycin (P/S; Life Technologies). SV40 immortalized human cardiomyocytes were cultured in Prigrow I medium (ABMGood) supplemented with 10% FBS (ABMGood) and 1% P/S. HepG2 cells were maintained in DMEM (Thermo-Fisher) supplemented with 10% FBS and 1% P/S. HL-1 and SV40 cardiomyocytes were grown on fibronectin/gelatin and applied extracellular matrix (ABMGood); whereas HepG2 cells did not require a matrix to grow. Cell lines were sustained in a humidified incubator supplemented with 5% CO2 at 37°C. All cultures were found to be negative for mycoplasma using a commercial PCR kit.

Hypoxia treatment of cells

A day prior to hypoxic incubation, routinely cultured cells were trypsinized and seeded into 6-wells plates at 1 million cells/well density. Medium was replaced with culture medium (without FBS) supplemented with either 100 μM pimonidazole or EF5. Cells were then transferred into InvivO2 chambers (Baker Ruskinn) and incubated at 37°C for 0, 1, 2, 3 or 6 hours. Composition of the atmosphere in the incubator consisted of 5% H_2, 5% CO_2, 0.2% O_2, and residual N_2. During each experiment, negative controls were incubated concurrently under normoxic conditions. After treatment, cells were directly put onto ice and either (1) lysed for Western blotting or (2) trypsinized for flow cytometry.

Flow cytometry analysis

Flow cytometry analysis was carried out as described by Koch et al. and Jankovic et al. [5,6]. Briefly, trypsinized cells were washed with cold PBS twice and fixed in 70% ethanol at 4°C for 30 minutes. In case of pimonidazole staining, cells were incubated in a block/stain-mixture at room temperature (RT) for 1 hour. This mixture consisted of 4% mouse serum (Dako), 0.1% Triton X-100 and 0.5 μg/mL anti-Pimo-FITC antibody. For EF5, cells

were blocked in a solution containing 0.3% PBS-Tween, 1.5% bovine serum albumin (BSA), 20% non-fat dry milk (Bio-Rad) and 5% mouse serum at 4°C overnight. After blocking, cells were washed once with PBS and stained with 10 µg/mL ELK3-51 antibodies at 4°C for 6 hours. Samples were analyzed using a FACSCanto II system (BD Biosciences). Fluorescein isothiocyanate (FITC) labels were excited with a 488-nm, air-cooled, solid-state argon ion laser and detected through a 530 ± 30 nm band pass (BP) filter. Sample analysis was completed when 10,000 events were counted and data were analyzed using FACSDiva™ software (BD Biosciences).

Cell lysis and Western blotting

Lysis of cells was carried out as described previously with slight modifications [7, 8]. SV40 and HL-1 media were collected and cells were washed twice with ice cold PBS before being lysed with freshly prepared CHAPS buffer (50 mM HEPES, pH 7.4, 150 mM KCl and 1% CHAPS) at 4°C for 2.5 hours. Lysates and media were centrifuged at 10,000g for 5 minutes and supernatant was collected. HepG2 cells were washed twice with ice cold PBS and lysed with freshly prepared lysis buffer (10 mM PBS, pH 7.4, 125 mM NaCl, 36 mM lithium dodecyl sulfate, 24 mM sodium deoxycholate, and 1% Triton X-100) supplemented with protease inhibitor cocktail and 1 mM PMSF. Cells were lysed at 4°C for 30 minutes, pulled through a 25-G needle 7-10 times, and finally centrifuged at 10,000g for 5 minutes.

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) and Western blotting was carried out as described previously [9, 10]. Briefly, SDS-PAGE was performed with 12% criterion XT precast gels (Bio-Rad) at 200V for 45 minutes. Proteins on the gels were then transferred to 0.45 µm nitrocellulose membranes at 100V for 1 hour. Application of antibodies was performed using the SNAP i.d.® 2.0 protein detection system (Merck-Millipore) according to the manufacturer's instructions. Both antibodies were incubated for 10 minutes with a washing step in between. The primary antibody was either mouse anti-Pimo unconjugated (0.1 µg/mL) or anti-EF5 (5 µg/mL). The secondary antibody was rabbit anti-FITC conjugated with HRP (1:10.000, Hypoxyprobe) for pimonidazole and goat-anti-Mouse conjugated with HRP (3:10.000, D0314) for EF5. After primary and secondary antibody application, the blots were incubated for 3 minutes in Supersignal West Femto chemiluminescent substrate (Thermo-Fisher) before detection with a Chemidoc XRS+ system (Bio-Rad). Images were processed with Quantity One (Bio-Rad) and ImageLab (version 6.0.1, Bio-Rad).

Synthesis of nitroimidazole-protein adducts in vitro

In order to synthesize protein-NI adducts in vitro we needed to derive identical nitroimidazole reaction products as found in vivo. In the current study we used (1) chemical Zinc/ammonium chloride reduction [11-14] and (2) radiochemical reduction [15, 16] to obtain these products. Besides pimonidazole and EF5 -which were employed in cell

studies- a fluorescent 2-nitroimidazole was in-house synthesized (DNI). Zinc reduction was carried out as described previously with minor modifications [11-14]. Briefly, a solution of 5 mM 2-nitroimidazole containing 0.5 mM ammonium chloride was de-aerated on ice by bubbling with a slow stream of ultrapure nitrogen for 30 minutes. Next, zinc dust (1 mg/mL) was added with continued nitrogen bubbling. Aliquots of the suspension were filtered after 20 minutes of incubation with zinc. These aliquots were added 1:1 to 1 a 0.01-1% protein solution in phosphate buffer (0.05 M; pH 6.8). To examine which reaction products are formed as a result of the Zinc reduction, we employed direct-infusion mass spectrometry at different time intervals. Radiochemical reduction was carried out as described previously [15, 16]. Briefly, proteins were dialyzed against three changes of phosphate buffered saline (pH 7.0) at 4°C, and re-reconstituted in formate buffer (100 mM; pH 5.5) at a concentration of 5 mg/mL protein. This solution was deoxygenated and irradiated to 720 Gy. Subsequently, 10 mM nitroimidazole, either pimonidazole or EF5, was added to the solution at a ratio of 1:20. The mixture was then deoxygenated and irradiated to 2000 Gy.

In-house synthesis of 1-dansylaminopropyl-2-nitroimidazole (DNI)

A fluorescent 2-nitroimidazole derivative, 1-dansylaminopropyl-2-nitroimidazole (DNI), was synthesized based on reported procedures [17-19]. Briefly, a suspension of 2-nitroimidazole (100 mg, 0.88 mmol) and potassium carbonate (366 mg, 2.64 mmol) in acetonitrile (ACN; 1 mL) was added methyl bromoacetate (270 mg, 1.76 mmol) at RT. The resultant reaction mixture was refluxed for 2 hours, cooled, diluted with ethyl acetate and washed with water. The organic layer was dried over anhydrous sodium sulfate and evaporated under reduced pressure to obtain methyl (2-nitro-1H-imidazol-1-yl)acetate (147 mg, 90%). An ice cooled solution of methyl (2-nitro-1H-imidazol-1-yl)acetate (147 mg, 0.8 mmol) in methanol (1.5 mL) was added to sodium hydroxide (63 mg, 1.6 mmol). The resultant reaction mixture was stirred at room temperature for 2-3 hours, subsequently neutralized with diluted HCl and extracted to ethyl acetate. The organic layer was dried over anhydrous sodium sulfate and evaporated under reduced pressure to obtain (2-nitro-1H-imidazol-1-yl)acetic acid (30 mg, 22 %). A solution of (2-nitro-1H-imidazol-1-yl)acetic acid (30 mg, 0.18 mmol) and N-hydroxysuccinimide (NHS) (24 mg, 0.21 mmol) in ACN (200 µL) was added a solution of N,N'-Dicyclohexylcarbodiimide (DCC) (47 mg, 0.23 mmol) in ACN (100 µL) at RT and stirred for 30 minutes. The resultant reaction mixture was filtered to remove dicyclohexylurea (DCU) by product. The mother liquor layer was added to a solution of dansyl amine[3] (54 mg, 0.18 mmol) and N,N-diisopropylethyl amine (23 mg, 0.18 mmol) at RT and stirred for 3 hours. The reaction mixture was evaporated to dryness and purified by silica gel column chromatography (5% methanol in chloroform) to obtain DNI as a yellow oil (40 mg, 50 %). ^{1}H NMR (400 MHz, Chloroform-d) d ppm: 1.51-1.61 (m, 2 H) 2.83 - 2.98 (m, 8 H) 3.22-3.32 (m, 2 H) 5.01 (s, 2 H) 5.97 (br. s., 1 H) 7.05 (br. s., 1 H) 7.08 - 7.20 (m, 2 H) 7.45 - 7.60 (m, 2 H) 8.15 (d, J=7.30 Hz, 1 H) 8.22 (d, J=8.56 Hz, 1 H) 8.52 (d, J=8.56 Hz, 1 H). Mass spectrometry, [M+Na^{+}]: 483.1415 Da.

Direct-infusion electrospray ionization mass spectrometry

To examine the products formed in our zinc reduction, we performed direct infusion analysis with aliquots taken at 5, 10, 15, 20, 25 and 30 minutes during the reduction. Aliquots were 1:1 dissolved in buffer (water: acetonitrile: 0.1% formic acid, 49.9:50:0.1, by vol.) and were analyzed by electrospray ionization mass spectrometry (ESI-MS) in positive ionization mode on a Quadrupole Time-of-Flight (QTOF) Micro mass spectrometer (Waters), using the direct infusion through a syringe with a constant flow rate of 10 µl/min, as previously described [20]. The QTOF mass spectrometer (micrOTOF, Bruker Daltonics) was operated with an ESI source for the direct-infusion method development. The following settings were applied: capillary voltage of -4.5kV, end plate offset of -500V, capillary exit of 250.0V, mass range of m/z 1000 to 8000, dry gas of 3.0 L min^{-1}, and drying temperature of 200°C.

Characterization of protein-nitroimidazole complexes

In order to increase our understanding of the protein-NI binding, we adapted several modifications to our original experiments. First, we compared the nitroimidazole binding in a set of four proteins with comprising at least 1 cysteine in their amino acid sequence; albumin, lysozyme, cTnI and transferrin, to two proteins without any cysteines in their sequence; skeletal troponin T (skTnT) and β-casein (see supplemental information C). We modified each of the proteins as previously described and assessed the nitroimidazole binding by Western blotting. Second, we chemically modified proteins by (1) reducing disulfides using tris-2 carboxyethylphosphine (TCEP) to increase their cysteine "content" and (2) alkylating cysteines using 2-iodoacetamide (IAA) to "cap" their cysteines. Reduction was carried out using 10 mM TCEP at 37°C for 30 minutes, and alkylation was optionally performed after reduction using 10, 50 or 100 mM IAA for 1 hour. Third, we mutated two cysteine to serine amino acids in cTnI. Wildtype and mutated cTnI (cTnI-C80S), with amino acid replacements C80S and C96S, were produced as previously described [21, 22]. Wildtype cTnI and mutated cTnI-C80S were expressed in Escherichia coli M15 (pREP4) (Qiagen), which were transformed with pQE-1 (GenScript) containing cDNA. Bacteria were harvested and lysed by sonication. Cell debris was removed by centrifugation. His-tagged proteins were isolated from supernatant by chromatography using nickel columns (GE Healthcare) and an imidazole gradient. Purified proteins were evaluated for protein content (Bradford assay) and purity (Coomassie staining).

Identification of nitroimidazole binding site by liquid chromatography-mass spectrometry

For a subset of proteins we performed liquid chromatography coupled with mass spectrometry (LC-MS) to unravel the exact binding site of nitroimidazoles to proteins. Proteins were modified using zinc or radiochemical reduction methods described previously. Protein-nitroimidazole adducts were then subjected to acetonitrile precipitation (ACN) by the addition of 400 µL of ice-cold (−20°C) ACN to 100 µL of protein-nitroimidazole mixture, followed by incubation at −20°C overnight. Samples were centrifuged at 12.000g

at 4 °C for 30 min, and the supernatant was decanted. Pellets were washed in 300 µL of ice-cold ACN, air dried, and redissolved in 50 µL of ammonium bicarbonate (AMBIC; 50 mmol/L, pH 8). Precipitated protein-nitroimidazole adducts were reduced with 20 mM dithiothreitol (DTT) at RT for 45 minutes, followed by alkylation in 40 mM IAA at RT in the dark for 45 minutes. Alkylation was terminated by the addition of 20 mM DTT. IAA and DTT solutions were dissolved in AMBIC. Proteins were then digested by a 1 µg/µL trypsin/lysC solution (Promega) at 37°C for 16 – 18 hours. TFA was added to the extracted peptides to a final concentration of 0.5% before LC–MS/MS analysis.

LC-MS/MS measurements were performed on a Q Exactive hybrid quadrupole-Orbitrap mass spectrometer, connected to an ultra-high-performance liquid chromatography (UHPLC) Dionex Ultimate 3000 (both by Thermo Fisher Scientific). Solvent A consisted of 0.1% formic acid in 100% water; solvent B consisted of 0.08% formic acid, 20% water, and 80% ACN. Peptides were first trapped on an Acclaim PepMap 100, 100 µm × 2 cm, C18, 5 µm, 100 A trap column in 0.1% TFA, 2% ACN, and 98% water. Next, peptides were separated on an Acclaim PepMap RSLC, 75 µm × 15 cm, C18, 2 µm, 100 A analytical column by a 180-min gradient of 4% to 45% solvent B. Washout was performed by a 5-min gradient of 55% to 90% solvent B and a constant flow of 90% solvent B for 5 min at 300 nL/min. Full MS1 scans were acquired in the Orbitrap with a width of 2.0Th at a resolution of 70000 full width at half maximum at 200 m/z, automated gain control (AGC) of 1e6, and a maximum injection time of 100ms. As a second scan event, a maximum of 10 most intense precursor ions within each full MS1 scan were selected for higher-energy collisional dissociation with an isolation window of 2.0 Th and a normalized collision energy of 30. Product ions were detected in the range of 250 to 1500 m/z at a resolution of 17500 full width at half maximum at 200 m/z, automated gain control target of 1e5, maximum injection time of 200ms, and a dynamic exclusion window of 30s.

All MS/MS spectra acquired were analyzed using the PEAKS software suite [23] We searched raw MS/MS spectra against databases containing our target proteins (i.e. troponin I, lysozyme, and albumin) concatenated with common lab contaminant proteins from the cRAP database. Search parameters included a parent mass error tolerance of 10.0 ppm, fragment mass error tolerance of 0.5 Da, enzyme trypsin with max missed cleavages of 2 and nonspecific trypsin cleavage, and static modifications of methionine (+57.02 Da) and oxidation (+16.00 Da). During our initial experiments, we searched for all theoretical products from a reduction reaction of a 2-nitroimidazole, but in our final analysis we limited ourselves to hydroxylamine- and amine- derivatives of a 2-nitroimidazole (see supplemental Table 2 for a complete list of m/z values). We selected all spectra with a FDR of ≥0.5% that had a high confidence peaks peptide score (≥20 –logP), a minimum of three consecutive b- or y-type peptide fragment ions and at least 2 peptide fragment ions surrounding the cysteine modification.

Results

Nitroimidazole binding in cellular systems is oxygen- and time-dependent

To verify the hypoxia-dependent binding mechanism of 2-nitroimidazoles, we first evaluated this in a cellular model system of hypoxia. Three different cells (SV40, HepG2 and HL-1 cells) were supplemented with 100 μM of 2-nitroimidazole (pimonidazole or EF5), deprived from oxygen for 0 to 6 hours and analyzed using either flow cytometry (Figure 1A) or Western blotting (Figure 1B). Flow cytometry (Figure 1A and 1C) and Western blotting (Figure 1B and 1D) reveal a strong hypoxia- and time-dependent signal for both pimonidazole and EF5. Notably, nitroimidazoles possessed distinctive binding characteristics; not only qualitatively, as reflected by different band patterns (Figure 1B), but also quantitatively, as observed both for FACS and WB (Figure 1C and 1D). These results further reveal that binding of nitroimidazoles was different between each of the cells.

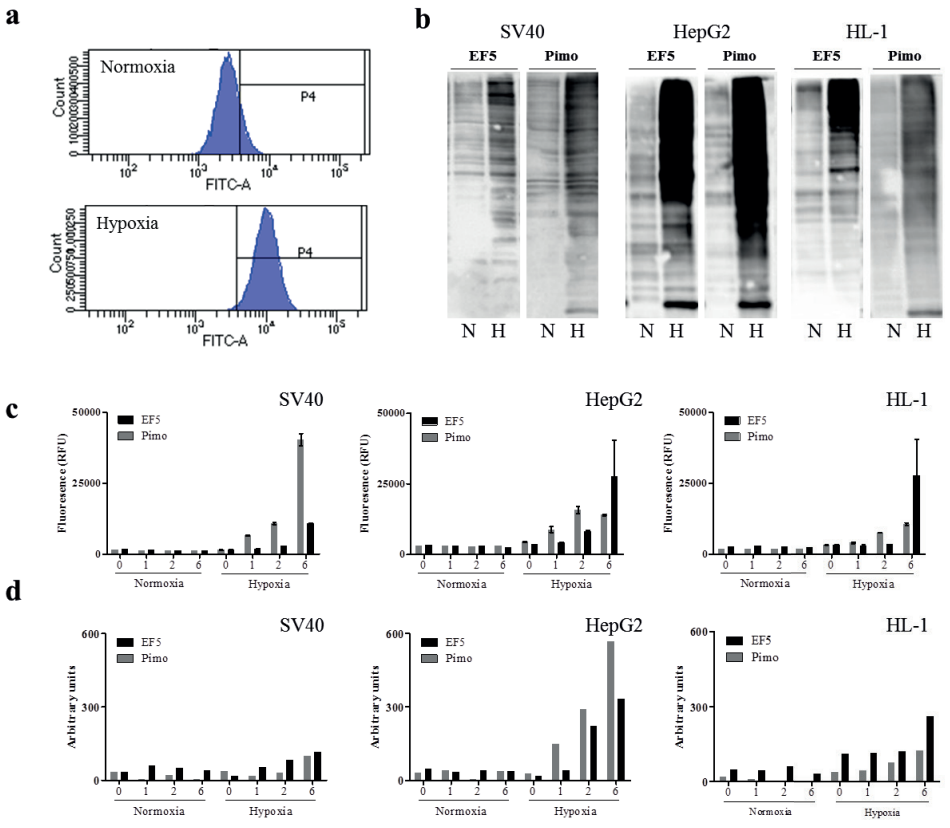


Figure 1. Nitroimidazole binding in cells is oxygen- and time-dependent. Cells were deprived from oxygen for 0 to 6 hours and analyzed using either flow cytometry (panel A) or Western blotting (B). Pimo and EF5 were observed to bind exclusively under hypoxic conditions as shown by flow cytometry (panel C) and Western blotting (panel D). Binding characteristics of nitroimidazole were different quantitatively, as reflected in different binding strengths, and qualitatively, as observed in the distinct band patterns in the Western blots (panel B). Additionally, binding characteristics were observed to be different in cell lines.

Synthesis and characterization of protein-nitroimidazole complexes

We next aimed to synthesize and examine protein-nitroimidazole adducts using in vitro methods. We employed a zinc/ammonium chloride reduction method that is capable of generating free electrons to reduce our 2-nitroimidazole compounds. Mass spectrometry analysis of initial experiments using this reduction methods confirms that it produces the desired nitroimidazole derivatives (see Supplemental Figure 1). Next, we evaluated the ability of a set of proteins to bind these reactive 2-nitroimidazoles (DNI, pimonidazole and EF5) by gel electrophoresis coupled with UV imaging (DNI) or Western Blotting (pimonidazole and EF5) (Figure 2A). Gel electrophoresis results reveal that 2-nitroimidazoles adducts exclusively formed when 2-nitroimidazoles were reduced. Additionally, we observe differences between the 2-nitroimidazoles. In contrast to proteins with a cysteine in their amino acid sequence, protein-nitroimidazole adducts were not detected on proteins without a cysteine site in their sequence (skTnT and β-casein, Figure 2A). To verify our initial results, we employed a second reduction method based on the generation of free electrons by radiation. The generated reactive 2-nitroimidazoles were also capable of binding to proteins (Figure 2B).

To further characterize the protein-nitroimidazole complexes, we adapted several modifications to our original experiments. Proteins were chemically modified by either reducing their disulfide bridges to free cysteines (Figure 2C, upper) or by alkylating their cysteines using iodoacetamide (Figure 2C, lower). These modified proteins were then bound to reactive DNI and analyzed using in-gel fluorescence. Results reveal an increased DNI binding in reduced proteins compared to native ones. Interestingly, DNI binding was diminished after alkylating the cysteine residues of albumin. In a second experiment, we mutated the two cysteines of troponin I (Cys80 and Cys96) to serines and analyzed the binding of DNI, pimonidazole and EF5 with wildtype and the mutated protein. Gel electrophoresis reveals that that wildtype cTnI was consistently able to bind nitroimidazole, which was not in the case in mutated cTnI (Figure 3D).

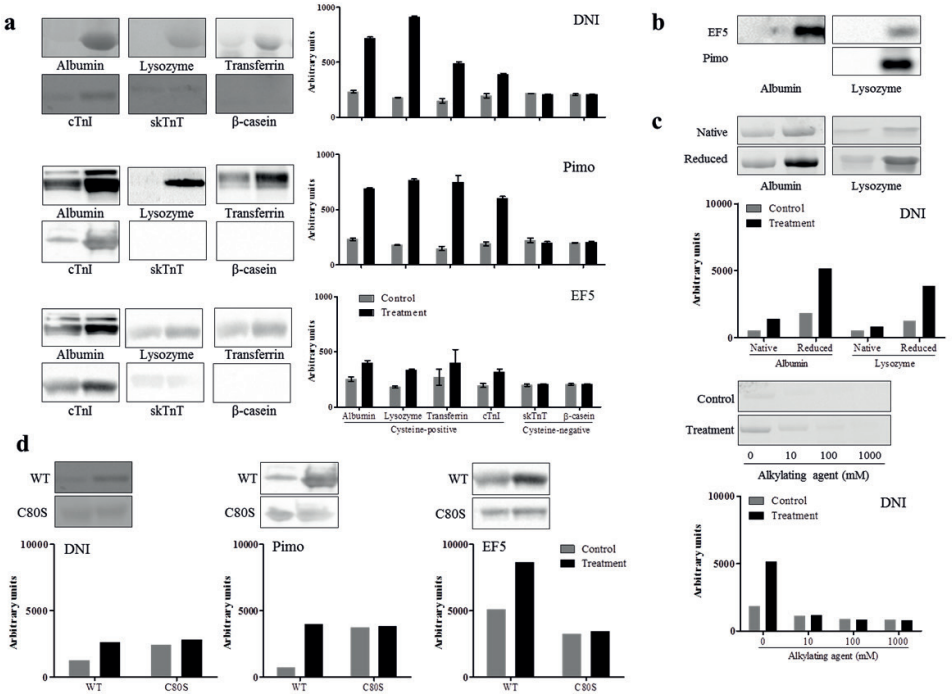


Figure 2. Synthesis and characterization of nitroimidazole-protein complexes. (A) Reaction products of pimo, EF5 and DNI were produced using a zinc-based reduction method, subsequently added to a set of 6 proteins (albumin, lysozyme, transferrin, cTnI, skTnT and β-casein) and detected using gel electrophoresis coupled with UV-imaging or Western blotting. (B) Pimo and EF5 were reduced using a radiochemical reduction method and detected as described in A. (C) Zinc-reduced DNI was added to chemically reduced albumin or lysozyme (upper panel) or chemically alkylated albumin (lower panel). (D) Zinc-reduced DNI was added to either wildtype (WT) or mutated (C80S) cTnI. The mutated C80S variant of cTnI had two cysteines substituted for serines.

Identification of nitroimidazole binding site in proteins

To identify which nitroimidazole derivative and amino acid are involved in the protein-nitroimidazole adduct formation, we performed mass spectrometry analysis in a subset of proteins. Troponin I, albumin and lysozyme were modified with either pimonidazole, EF5 and DNI using the zinc and radiochemical reduction methods. These adducts were digested to peptides using trypsin and analyzed with liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS), allowing us identify the derivative and amino acid site involved in the adduct formation. We consistently found that a cysteine site was modified by the amine-derivative of the nitroimidazole. For example, we found mass shifts of 270.05 Da that correspond to amine-EF5 adducts occurred on peptide 287YIC(amine-EF5)DNQDTISSK298 from albumin (Figure 3A, left) and peptide 80C(amine-EF5)QPLELAGLGFAELQDLCR98 from troponin (Figure 3B, right). Troponin was modified exclusively at Cys80 (Figure 3B and supplemental Table 2), whilst albumin and lysozyme were found to have numerous cysteine sites modified (Figure 3B and supplemental Table 3 and 4). Although nitroimidazoles often targeted identical cysteines in albumin and lysozyme, differences sites were targeted by nitroimidazoles. We validated our findings by analyzing protein-nitroimidazole adducts derived by zinc reduction (see supplemental Table 2-4). These adducts were found on similar sites as observed with radiochemical reduction although often less frequent and intense. Based on these results we propose the mechanism of protein-nitroimidazole formation as depicted in Figure 3C. Herein, the hydroxylamine derivative of a nitroimidazole compounds initiates the sulfide binding between the C4 or C5 of the imidazole ring with a thiol group of the cysteine amino acid. During this binding, the hydroxylamine derivative is stabilized by the conversion to an amine group.

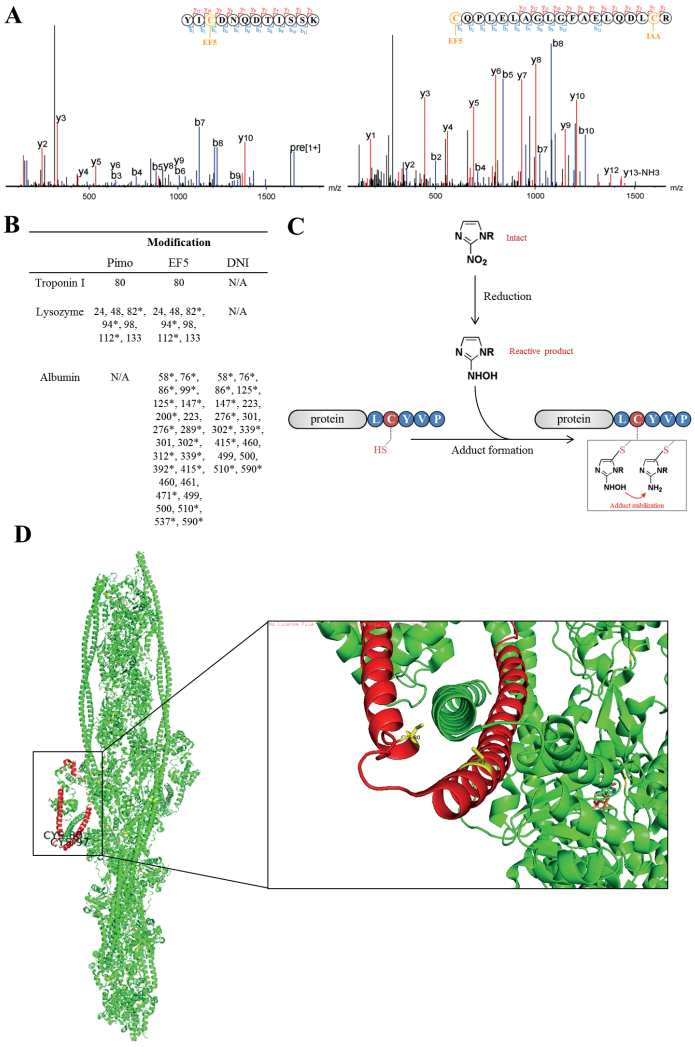


Figure 3. Mass spectrometric identification of nitroimidazole binding compound and site. (A) Illustrative MS/MS spectra of peptide 286YIC(EF5)DNQDTISSK297 originating from albumin (left) and 80C(EF5)QPLELAGLGFAELQDLCR98 from troponin I (right). MS2 spectra were analyzed for ions corresponding to the additional mass of +270.05 Da (amino-EF5). The fragmentation pattern for these peptides shows the mass shift of the amine-EF5 modification on Cys289 and Cys80, respectively. (B) Overview of detected nitroimidazole modifications on troponin I, lysozyme and albumin. Residues marked with stars were found with Zinc and radiochemical reduction. (C) Proposed reaction mechanism of nitroimidazole binding to proteins. Through a series of reactions, the nitro group in the nitroimidazole is reduced to a hydroxylamine. The hydroxylamine binds to proteins through the sulfur originating from a cysteine; whilst binding occurs the hydroxylamine is further reduced to an amine, ultimately resulting in a stable 2-nitroimidazole adduct on proteins. (D) Overview of the 3D structure of the cardiac troponin complex with troponine I colored red. The targeted cysteine molecules are shown in yellow with their corresponding molecules bound.

Discussion

The current study characterizes the binding of 2-nitroimidazoles to proteins during hypoxia. We observed nitroimidazole-protein adduct formation in hypoxic cells, which through examination of in vitro synthesized protein-nitroimidazole adducts was shown to occur through a sulfide bond with cysteine amino acids. Our results provide a basis for assay development specifically detecting hypoxia-dependent biomarker release.

2-Nitroimidazoles are known to bind to cellular macromolecules, including proteins[5,31,32], DNA[33-35] and low-level metabolites such as glutathione[8,36-39]. The configuration of protein-nitroimidazole adducts has been studied for over four decades[5,40,41], showing that the hydroxylamine derivative most likely binds to protein thiols[5,19,42]. Nonetheless, insight in the exact configuration of the binding site and derivative was still lacking. Using synthesized protein-nitroimidazole adducts, we were able to identify cysteine as the anchor for the hydroxylamine derivative to bind through a sulfide bond with either the C4 or C5 of the imidazole ring (Figure 3C). During this binding, the hydroxylamine is stabilized by conversion to the amine. Hence, we consistently found amine derivatives of pimonidazole, EF5 and DNI on various proteins including cardiac troponin I, lysozyme and albumin.

Radioactive tracers of 2-nitroimidazoles are successfully applied for imaging tumor hypoxia using position emission tomography (PET) and single photon emission computed tomography (SPECT)[43-45]. Translation of this biological concept to an application that would enable hypoxia-specific biomarker detection has not been previously explored.

This study provides the scientific basis for such endeavor. Possible applications of hypoxia specific biomarker assays vary from hypoxia research questions, to clinical diagnostic settings where discrimination between hypoxic and non-hypoxic cell death is critical to guide therapy. Perhaps convenient for translation beyond applications in cells and animals, is the fact that the 2-nitroimidazoles pimonidazole and EF5 are FDA approved, and its administration to patients is already common practice in nuclear medicine. The textbook example of a diagnostic trajectory that would dramatically improve from rapid discrimination between hypoxic and non-hypoxic cell death is the use of cardiac troponin in patients presenting with chest pain. Cardiac troponin is highly specific to myocardial tissue, and has become the biochemical gold standard for the diagnosis of acute myocardial infarction[46,47]. However, despite troponin's excellent sensitivity and cardiac tissue specificity, the diagnosis of acute myocardial infarction remains often challenging[48,49]. This diagnostic complexity stems from the inability to discriminate between troponin elevations due to hypoxic cardiac injury, and troponin elevations from other causes of cardiac cell death, such as acute heart failure, myocarditis, tachyarrhythmia, and hypertensive emergency, all clinical entities that are characterized by an overlapping spectrum of clinical presentations.

Proof-of-concept experiments in the present study demonstrate that in vitro synthesized cardiac troponin I-nitroimidazole adducts, can be specifically detected with LC-MS, and provide a basis for developing an hypoxia-dependent troponin assay exploiting the Cys80-nitroimidazole adduct on cardiac troponin I.

References

1. Claycomb, W.C., et al., HL-1 cells: a cardiac muscle cell line that contracts and retains phenotypic characteristics of the adult cardiomyocyte. Proc Natl Acad Sci U S A, 1998. 95(6): p. 2979-84.
2. Durham, K.K., et al., HDL protects against doxorubicin-induced cardiotoxicity in a scavenger receptor class B type 1-, PI3K-, and Akt-dependent manner. Am J Physiol Heart Circ Physiol, 2018. 314(1): p. H31-H44.
3. Kono, K., et al., Development of selective cytotoxic viral vectors for concentration of undifferentiated cells in cardiomyocytes derived from human induced pluripotent stem cells. Sci Rep, 2019. 9(1): p. 3630.
4. Schoonen, W.G., et al., Cytotoxic effects of 110 reference compounds on HepG2 cells and for 60 compounds on HeLa, ECC-1 and CHO cells. II mechanistic assays on NAD(P)H, ATP and DNA contents. Toxicol In Vitro, 2005. 19(4): p. 491-503.
5. Koch, C.J., Importance of antibody concentration in the assessment of cellular hypoxia by flow cytometry: EF5 and pimonidazole. Radiat Res, 2008. 169(6): p. 677-88.
6. Jankovic, B., et al., Comparison between pimonidazole binding, oxygen electrode measurements, and expression of endogenous hypoxia markers in cancer of the uterine cervix. Cytometry B Clin Cytom, 2006. 70(2): p. 45-55.
7. Streng, A.S., et al., Cardiac troponin in ischemic cardiomyocytes: intracellular decrease before onset of cell death. Exp Mol Pathol, 2014. 96(3): p. 339-45.
8. Meex, S.J., et al., Huh-7 or HepG2 cells: which is the better model for studying human apolipoprotein-B100 assembly and secretion? J Lipid Res, 2011. 52(1): p. 152-8.
9. Cardinaels, E.P., et al., Time-dependent degradation pattern of cardiac troponin T following myocardial infarction. Clin Chem, 2013. 59(7): p. 1083-90.
10. Michielsen, E.C., et al., Highly sensitive immunoprecipitation method for extracting and concentrating low-abundance proteins from human serum. Clin Chem, 2005. 51(1): p. 222-4.
11. Varghese, A.J. and G.F. Whitmore, Properties of 2-hydroxylaminoimidazoles and their implications for the biological effects of 2-nitroimidazoles. Chem Biol Interact, 1985. 56(2-3): p. 269-87.
12. Bolton, J.L. and R.A. McClelland, Kinetics and mechanism of the decomposition in aqueous solutions of 2-(hydroxyamino)imidazoles. Journal of the American Chemical Society, 1989. 111(21): p. 8172-8181.
13. McClelland, R.A., R. Panicucci, and A.M. Rauth, Products of reductions of 2-nitroimidazoles. Journal of the American Chemical Society, 1987. 109(14): p. 4308-4314.
14. Varghese, A.J. and G.F. Whitmore, Cellular and chemical reduction products of misonidazole. Chem Biol Interact, 1981. 36(2): p. 141-51.
15. Koch, C.J. and J.A. Raleigh, Radiolytic reduction of protein and nonprotein disulfides in the presence of formate: a chain reaction. Arch Biochem Biophys, 1991. 287(1): p. 75-84.
16. Raleigh, J.A. and C.J. Koch, Importance of thiols in the reductive binding of 2-nitroimidazoles to macromolecules. Biochem Pharmacol, 1990. 40(11): p. 2457-64.
17. Pavlik, C., et al., Synthesis and fluorescent characteristics of imidazole–indocyanine green conjugates. Dyes and Pigments, 2011. 89(1): p. 9-15.
18. Wei, H., et al., Design and Synthesis of Vandetanib Derivatives Containing Nitroimidazole Groups as Tyrosine Kinase Inhibitors in Normoxia and Hypoxia. Molecules, 2016. 21(12).
19. Bhoi, A.K., et al., Analyte interactions with a new ditopic dansylamide-nitrobenzoxadiazole dyad: a combined photophysical, NMR, and theoretical (DFT) study. J Phys Chem B, 2014. 118(33): p. 9926-37.
20. Aslebagh, R., et al., Mass spectrometry-based proteomics of oxidative stress: Identification of 4-hydroxy-2-nonenal (HNE) adducts of amino acids using lysozyme and bovine serum albumin as model proteins. Electrophoresis, 2016. 37(20): p. 2615-2623.
21. Stohr, R., et al., AnnexinA5-pHrodo: a new molecular probe for measuring efferocytosis. Sci Rep, 2018. 8(1): p. 17731.
22. Wildhagen, K.C., et al., Nonanticoagulant heparin prevents histone-mediated cytotoxicity in vitro and improves survival in sepsis. Blood, 2014. 123(7): p. 1098-101.

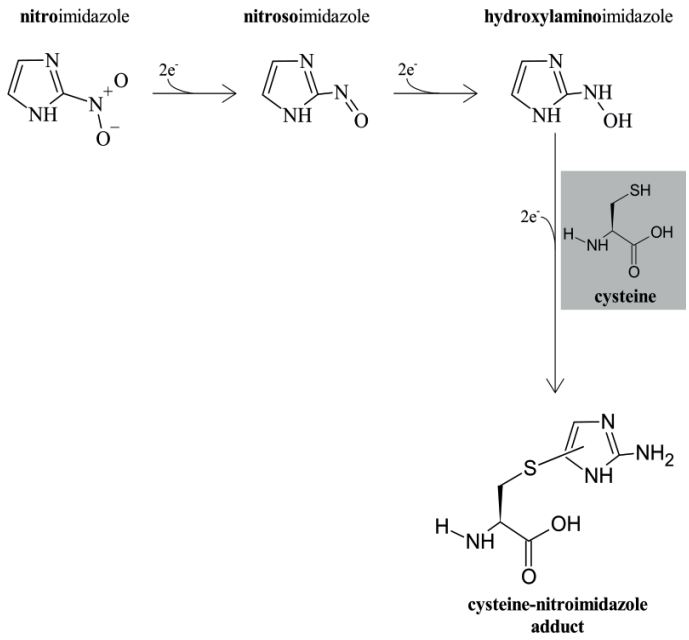
23. Zhang, J., et al., PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. Mol Cell Proteomics, 2012. 11(4): p. M111 010587.
24. Knox, R.J., R.C. Knight, and D.I. Edwards, Studies on the action of nitroimidazole drugs. The products of nitroimidazole reduction. Biochem Pharmacol, 1983. 32(14): p. 2149-56.
25. Kizaka-Kondoh, S. and H. Konse-Nagasawa, Significance of nitroimidazole compounds and hypoxia-inducible factor-1 for imaging tumor hypoxia. Cancer Sci, 2009. 100(8): p. 1366-73.
26. Rowley, D.A., et al., The effect of nitroheterocyclic drugs on DNA: an in vitro model of cytotoxicity. Biochem Pharmacol, 1979. 28(19): p. 3009-13.
27. Aslebagh, R., et al., Mass spectrometry-based proteomics of oxidative stress: Identification of 4-hydroxy-2-nonenal (HNE) adducts of amino acids using lysozyme and bovine serum albumin as model proteins. Electrophoresis, 2016. 37(20): p. 2615-2623.
28. Wada, T., et al., Effect of misonidazole on radiation-induced reduction of DNA bases in deaerated aqueous solution. J Radiat Res, 1984. 25(1): p. 99-110.
29. Mascini, N.E., et al., Mass Spectrometry Imaging of the Hypoxia Marker Pimonidazole in a Breast Tumor Model. Anal Chem, 2016. 88(6): p. 3107-14.
30. Masaki, Y., et al., FMISO accumulation in tumor is dependent on glutathione conjugation capacity in addition to hypoxic state. Ann Nucl Med, 2017. 31(8): p. 596-604.
31. Masaki, Y., et al., Imaging Mass Spectrometry Revealed the Accumulation Characteristics of the 2-Nitroimidazole-Based Agent "Pimonidazole" in Hypoxia. PLoS One, 2016. 11(8): p. e0161639.
32. Masaki, Y., et al., The accumulation mechanism of the hypoxia imaging probe "FMISO" by imaging mass spectrometry: possible involvement of low-molecular metabolites. Sci Rep, 2015. 5: p. 16802.
33. Varghese, A.J., Glutathione conjugates of misonidazole. Biochem Biophys Res Commun, 1983. 112(3): p. 1013-20.
34. Whitmore, G.F., S. Gulyas, and A.J. Varghese, Sensitizing and toxicity properties of misonidazole and its derivatives. Br J Cancer Suppl, 1978. 3: p. 115-9.
35. Wang, L., et al. Treasure hunt for peptides with undefined chemical modifications: Proteomics identification of differential albumin adducts of 2-nitroimidazole-indocyanine green in hypoxic tumor. J Mass Spectrom 2019 May 25; 2019/05/28:[Available from: https://www.ncbi.nlm.nih.gov/pubmed/31128078 https://onlinelibrary.wiley.com/doi/pdf/10.1002/jms.4376.
36. McClelland, R.A., et al., 2-Hydroxylaminoimidazoles--unstable intermediates in the reduction of 2-nitroimidazoles. Biochem Pharmacol, 1984. 33(2): p. 303-9.
37. Leitsch, D., et al., Nitroimidazole action in Entamoeba histolytica: a central role for thioredoxin reductase. PLoS Biol, 2007. 5(8): p. e211.
38. Camerini, S., et al., Proteomic and functional analyses reveal pleiotropic action of the anti-tumoral compound NBDHEX in Giardia duodenalis. Int J Parasitol Drugs Drug Resist, 2017. 7(2): p. 147-158.
39. Lalle, M., et al., The FAD-dependent glycerol-3-phosphate dehydrogenase of Giardia duodenalis: an unconventional enzyme that interacts with the g14-3-3 and it is a target of the antitumoral compound NBDHEX. Front Microbiol, 2015. 6: p. 544.
40. Ballinger, J.R., Imaging hypoxia in tumors. Semin Nucl Med, 2001. 31(4): p. 321-9.
41. Lopci, E., et al., PET radiopharmaceuticals for imaging of tumor hypoxia: a review of the evidence. Am J Nucl Med Mol Imaging, 2014. 4(4): p. 365-84.
42. Fleming, I.N., et al., Imaging tumour hypoxia with positron emission tomography. Br J Cancer, 2015. 112(2): p. 238-50.
43. Lee, S.T. and A.M. Scott, Hypoxia positron emission tomography imaging with 18f-fluoromisonidazole. Semin Nucl Med, 2007. 37(6): p. 451-61.
44. Nie, X., et al., Imaging of hypoxia in mouse atherosclerotic plaques with (64)Cu-ATSM. Nucl Med Biol, 2016. 43(9): p. 534-542.
45. Takahashi, N., et al., Copper-62 ATSM as a hypoxic tissue tracer in myocardial ischemia. Ann Nucl Med, 2001. 15(3): p. 293-6.

46. Pell, V.R., et al., PET Imaging of Cardiac Hypoxia: Hitting Hypoxia Where It Hurts. Curr Cardiovasc Imaging Rep, 2018. 11(3): p. 7.
47. Neumann, J.T., et al., Application of High-Sensitivity Troponin in Suspected Myocardial Infarction. N Engl J Med, 2019. 380(26): p. 2529-2540.
48. Westermann, D., et al., High-sensitivity assays for troponin in patients with cardiac disease. Nat Rev Cardiol, 2017. 14(8): p. 472-483.
49. Glatz, J. and R. Renneberg, Added value of H-FABP as plasma biomarker for the early evaluation of suspected acute coronary syndrome. Clinical Lipidology, 2014. 9: p. 205-220.
50. Glatz, J.F., et al., Release of fatty acid-binding protein from isolated rat heart subjected to ischemia and reperfusion or to the calcium paradox. Biochim Biophys Acta, 1988. 961(1): p. 148-52.
51. Baker, J.O., et al., Cardiac myosin-binding protein C: a potential early biomarker of myocardial injury. Basic Res Cardiol, 2015. 110(3): p. 23.
52. Kaier, T.E., et al., Direct Comparison of Cardiac Myosin-Binding Protein C With Cardiac Troponins for the Early Diagnosis of Acute Myocardial Infarction. Circulation, 2017. 136(16): p. 1495-1508.
53. Stroka, D.M., et al., HIF-1 is expressed in normoxic tissue and displays an organ-specific regulation under systemic hypoxia. FASEB J, 2001. 15(13): p. 2445-53.
54. Semenza, G.L., et al., Hypoxia response elements in the aldolase A, enolase 1, and lactate dehydrogenase A gene promoters contain essential binding sites for hypoxia-inducible factor 1. J Biol Chem, 1996. 271(51): p. 32529-37.
55. Bhardwaj, A., et al., A multicenter comparison of established and emerging cardiac biomarkers for the diagnostic evaluation of chest pain in the emergency department. Am Heart J, 2011. 162(2): p. 276-282 e1.
56. Kim, J.S., et al., Ischemia-modified albumin: is it a reliable diagnostic and prognostic marker for myocardial ischemia in real clinical practice? Cardiology, 2010. 116(2): p. 123-9.
57. Danne, O. and M. Mockel, Choline in acute coronary syndrome: an emerging biomarker with implications for the integrated assessment of plaque vulnerability. Expert Rev Mol Diagn, 2010. 10(2): p. 159-71.
58. Ohkawa, R., et al., Measurement of plasma choline in acute coronary syndrome: importance of suitable sampling conditions for this assay. Sci Rep, 2018. 8(1): p. 4725.
59. Body, R., et al., Choline for diagnosis and prognostication of acute coronary syndromes in the Emergency Department. Clin Chim Acta, 2009. 404(2): p. 89-94.
60. Bhagavan, N.V., et al., Utility of serum Fatty Acid concentrations as a marker for acute myocardial infarction and their potential role in the formation of ischemia-modified albumin: a pilot study. Clin Chem, 2009. 55(8): p. 1588-90.
61. Apple, F.S., A.M. Kleinfeld, and J. Adams, Unbound free fatty acid concentrations are increased in cardiac ischemia. Clinical Proteomics, 2004. 1(1): p. 41-44.
62. Lord, E.M., L. Harwell, and C.J. Koch, Detection of hypoxic cells by monoclonal antibody recognizing 2-nitroimidazole adducts. Cancer Res, 1993. 53(23): p. 5721-6.
63. DeGraff, W.G., et al., Evaluation of nitroimidazole hypoxic cell radiosensitizers in a human tumor cell line high in intracellular glutathione. Int J Radiat Oncol Biol Phys, 1989. 16(4): p. 1021-4.
64. Kaanders, J.H., et al., Pimonidazole binding and tumor vascularity predict for treatment outcome in head and neck cancer. Cancer Res, 2002. 62(23): p. 7066-74.
65. Russell, J., et al., Immunohistochemical detection of changes in tumor hypoxia. Int J Radiat Oncol Biol Phys, 2009. 73(4): p. 1177-86.

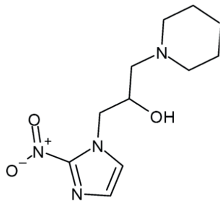
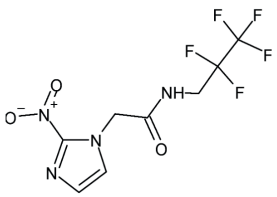
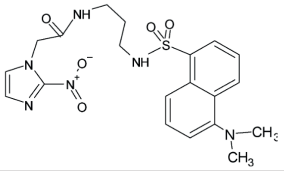
Supplemental material

Supplemental information

Supplemental information A. Proposed reaction mechanism of 2-nitroimidazoles with proteins. 2-nitroimidazoles have previously been described to bind the cellular macromolecules, including proteins [1-3], DNA [3, 4] and low-level metabolites such as glutathione [5, 6]. The proposed reaction mechanism of 2-nitroimidazoles with proteins is depicted in the figure underneath.



Supplemental information B. Structural information of 2-nitroimidazoles used in current study.
2-nitroimidazoles are a group of molecules characterized by their nitro group at the second position of the imidazole ring. Numerous 2-nitroimidazoles have been synthesized, studied and evaluated in basic and clinical research; refer to review articles for an in-depth examination of these topics [7, 8]. Structures, formulas and partition coefficients of the 2-nitroimidazoles used in the current study are described in the table underneath.

Common name	Formula	Avg. mass (mono-isotopic)	Structure	Partition coefficient[1]
Pimonidazole	$C_{11}H_{18}N_4O_3$	254.2861 (254.1379)		*P* = 8.5 (literature) log P = 1.00 (±0.58) (calculated)
EF5	$C_8H_7F_5N_4O_3$	302.1585 (302.0438)		log P = 1.1 (literature) log P = 1.35 (±0.83) (calculated)
DNI	$C_{20}H_{24}N_6O_5S$	460.5074 (460.1529)		log P = 1.1 (literature) log P = 1.88 (±0.55) (calculated)

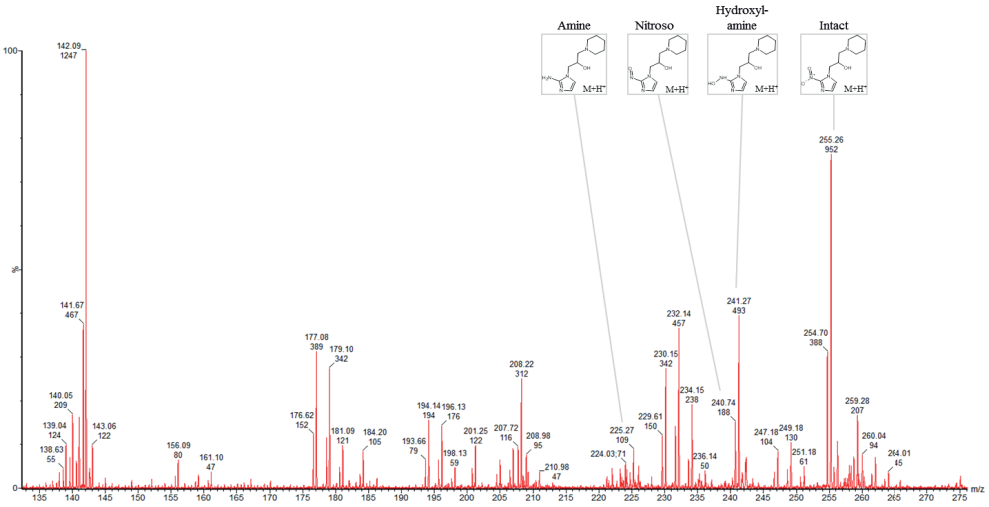
[1] Partition coefficient was retrieved from Kizaka-Kondoh et al. [7], calculations were performed by ChemSketch (version 2019.2.1, ACD/Labs).

Supplemental information C. Amino acid sequences of proteins used in the current study. List of protein sequences used in the current study with cysteine amino acids highlighted.

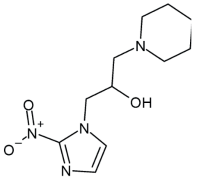
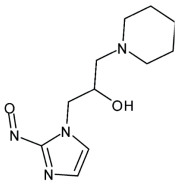
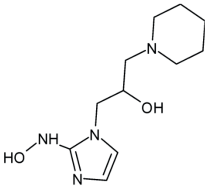
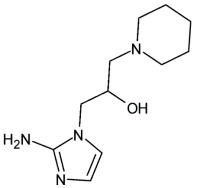
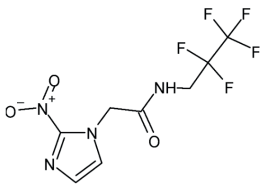
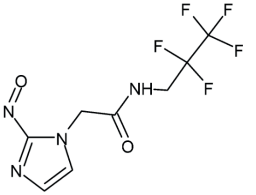
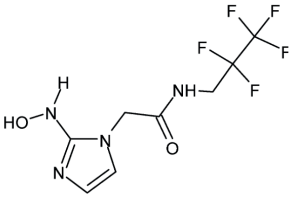
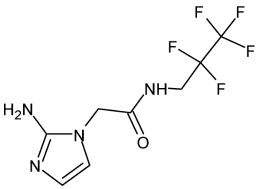
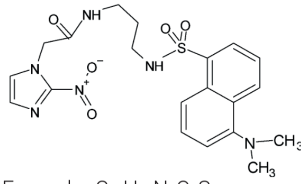
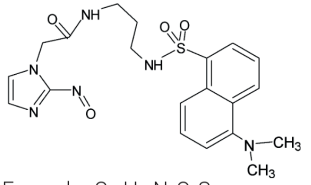
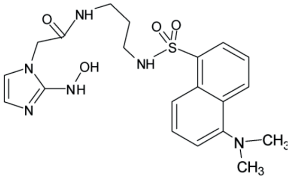
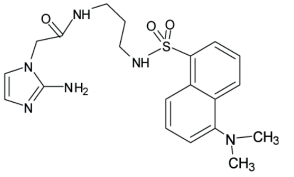
Protein	Sequence
Albumin	MKWVTFISLLLLFSSAYSRGVFRRDTHKSEIAHRFKDLGEEHFKGLVLI AFSQYLQQ**C**PFDEHVKLVNELTEFAKT**C**VADESHAG**C**EKSLHTLFGDEL**C**KVASL RETYGDMAD**CC**EKQEPERNE**C**FLSHKDDSPDLPKLKPDPNTL**C**DEFKADEKKFW GKYLYEIARRHPYFYAPELLYYANKYNGVFQE**CC**QAEDKGA**C**LLPKIETMREKVLAS SARQRLR**C**ASIQKFGERALKAWSVARLSQKFPKAEFVEVTKLVTDLTKVHKE**CC** HGDLLE**C**ADDRADLAKYI**C**DNQDTISSKLKE**CC**DKPLLEKSH**C**IAEVEKDAIPEN LPPLTADFAEDKDV**C**KNYQEAKDAFLGSFLYEYSRRHPEYAVSVLLRLAKEYEAT LEE**CC**AKDDPHA**C**YSTVFDKLKHLVDEPQNLIKQN**C**DQFEKLGEYGFQNALIVRY TRKVPQVSTPTLVEVSRSLGKVGTR**CC**TKPESERMP**C**TEDYLSLIL NRL**C**VLHEKTPVSEKVTK**CC**TESLVNRRP**C**FSALTPDETYVPKAFDEKLFTFHADI**C**T LPDTEKQIKKQTALVELLKHKPKATEEQLKTVMENFVAFVDK**CC**AADDKEA**C**FAVEGP KLVVSTQTALA
Troponin I	MADGSSDAAREPRPAPAPIRRRSSNYRAYATEPHAKKKSKISASRKLQLK TLLLQIAKQELEREAEERRGEKGRALSTR**C**QPLELAGLGFAELQDL**C**RQLHARVDK VDEERYDIEAKVTKNITEIADLTQKIFDLRGKFKRPTLRRVRISADAMMQALLGAR AKESLDLRAHLKQVKKEDTEKENREVGDWRKNIDALSGMEGRKKKFESL
Lysozyme	MRSLLILVL**C**FLPLAALGKVFGR**C**ELAAAMKRHGLDNYRGYSLGNWV**C**AAKFESN FNTQATNRNTDGSTDYGILQINSRWW**C**NDGRTPGSRNL**C**NIP**C**SALLSSDITASVN **C**AKKIVSDGNGMNAWVAWRNR**C**KGTDVQAWIRG**C**RL
Transferrin	MRLAVGALLV**C**AVLGL**C**LAVPDKTVRW**C**AVSEHEATK**C**QSFRDHMKSVIPSDG PSVA**C**VKKASYLD**C**IRAIAANEADAVTLDAGLVYDAYLAPNNLKPVVAEFYGSKED PQTFYYAVAVVKKDSGFQMNQLRGKKS**C**HTGLGRSAGWNIPIGLLY**C**DLPEPRK PLEKAVANFFSGS**C**AP**C**ADGTDFPQL**C**QL**C**PG**C**G**C**STLNQYFGYSGAFK**C**LKDGAG DVAFVKHSTIFENLANKADRDQYELL**C**LDNTRKPVDEYKD**C**HLAQVPSHTVVARSIG GKEDLIWELLNQAQEHFGKDKSKEFQLFSSPHGKDLLFKDSAHGFLKVPPRMDA KMYLGYEYVTAIRNLREGT**C**PEAPTDE**C**KPVKW**C**ALSHHERLK**C**DEWSVNSVGK IE**C**VSAETTED**C**IAKIMNGEADAMSLDGGFVYIAGK**C**GLVPVLAENYNKSDN**C**EDTP GAGYFAVAVVKKSASDLTWDNLKGKKS**C**HTAVGRTAGWNIPMGLLYNKINH**C**RF DEFFSEG**C**APGSKKDSSL**C**KL**C**MGSGLNL**C**EPNNKEGYYGYTGAFR**C**LVEKGDVAF VKHQTVPQNTGGKNPDPWAKNLNEKDYELL**C**LDGTRKPVEEYAN**C**HLARAPNHAV VTRKDKEA**C**VHKILRQQQHLFGSNVTD**C**SGNF**C**LFRSETKDLLFRDDTV**C**LAKLH DRNTYEKYLGEEYVKAVGNLRK**C**STSSLLEA**C**TFRRP
Skeletal troponin T (skTnT)	MSDEEVEQVEEQYEEEEEAQEEAAEVHEEVHEPEEVQEDTAEEDAEEEKPRP KLTAPKIPEGEKVDFDDIQKKRQNKDLMELQALIDSHFEARKKEEEELVALKER IEKRRAERAEQQRIRAEKERERQNRLAEEKARREEEDAKRRAEDDLKKKKALSSMGA NYSSYLAKADQKRGKKQTAREMKKKILAERRKPLNIDHLGEDKLRDKAKELWETLH QLEIDKFEFGEKLKRQKYDITTLRSRIDQAQKHSKKAGTPAKGKVGGRWK
β-Casein	RELEELNVPGEIVESLSSSEESITRINKKIEKFQSEEQQQTEDELQDKIHP FAQTQSLVYPFPGPIPNSLPQNIPPLTQTPVVVPPFLQPEVMGVSKVKEAMAPK HKEMPFPKYPVEPFTESQSLTLTDVENLHLPLPLLQSWMHQPHQPLPPTVMFP PQSVLSLSQSKVLPVPQKAVPYPQRDMPIQAFLLYQEPVLGPVRGPFPIIV

Supplemental Figures

Supplemental Figure 1. Direct infusion ESI-MS reaction products Zinc reduction.
In order to verify that the Zinc reduction method produces the reactive derivatives we are aiming for, we analyzed the reaction mixture using direct infusion ESI-MS. An illustrative example of the ESI-MS analysis of the zinc reduction mixture with pimonidazole is shown below. Through analysis of the mass spectra, each of the reaction products are observed as annotated in the figure below.



Supplemental table 1. Nitroimidazoles and their theoretical derivatives used in current study.

Compound	Derivatives			
	Intact	**Nitroso-**	**Hydroxylamine-**	**Amine-**
Pimonidazole	 Formula: $C_{11}H_{18}N_4O_3$ Avg. mass: 254.2861 M.I. mass: 254.1379	 Formula: $C_{11}H_{18}N_4O_2$ Avg. mass: 238.2867 M.I. mass: 238.1430	 Formula: $C_{11}H_{20}N_4O_2$ Avg. mass: 240.3026 M.I. mass: 240.1586	 Formula: $C_{11}H_{20}N_4O_1$ Avg. mass: 224.3032 M.I. mass: 224.1637
EF5	 Formula: $C_8H_7F_5N_4O_3$ Avg. mass: 302.1585 M.I. mass: 302.0438	 Formula: $C_8H_7F_5N_4O_2$ Avg. mass: 286.1591 M.I. mass: 286.0489	 Formula: $C_8H_9F_5N_4O_2$ Avg. mass: 288.1750 M.I. mass: 288.0646	 Formula: $C_8H_9F_5N_4O_1$ Avg. mass: 272.1756 M.I. mass: 272.0697
DNI	 Formula: $C_{20}H_{24}N_6O_5S$ Avg. mass: 460.5074 M.I. mass: 460.1529	 Formula: $C_{20}H_{24}N_6O_4S$ Avg. mass: 444.5080 M.I. mass: 444.1580	 Formula: $C_{20}H_{26}N_6O_4S$ Avg. mass: 446.5239 M.I. mass: 446.1736	 Formula: $C_{20}H_{26}N_6O_3S$ Avg. mass: 430.5245 M.I. mass: 430.1787

Supplemental table 2. Mass spectrometry variable modification searches.
In order to find all potential nitroimidazole derivatives that can bind to cysteine amino acids, we initially searched for all of them in our mass spectrometry searches. We hypothesized that during adduct formation there would be a double proton loss: one from the thiol (-SH) group, and one from the imidazole ring in the nitroimidazole compound. Hence, we can derive the mass shifts (m/z) from supplemental Table 1.

Compound	m/z shift for mass spectrometry searches			
	Intact	Nitroso-	Hydroxylamine-	Amine-
Pimonidazole	252.1222	236.1273	238.1430	222.1481
EF5	300.0282	284.0333	286.0489	270.0540
DNI	458.1372	442.1423	444.1580	428.1631

Supplemental table 3. Mass spectrometric identification of nitroimidazole adducts on troponin I.
List of modified peptides that were found for the trypsin digestion of troponin I adducts. # denotes a oxidation modification (+16.00 Da) and * denotes a carbamidomethylation (+57.02 Da). Z indicates zinc reduction; R indicates radiochemical reduction.

Peptide	Modifications	Method
^{80}CQPLELAGLGFAELQDLCR^{98}	Amine-EF5 (+270.05 m/z) on **Cys80** Amine-Pimo (+222.15 m/z) on **Cys80**	Z Z

Supplemental table 4. Mass spectrometric identification of nitroimidazole adducts on albumin.
List of modified peptides that were found for the trypsin digestion of albumin adducts. # denotes a oxidation modification (+16.00 Da) and * denotes a carbamidomethylation (+57.02 Da). Z indicates zinc reduction; R indicates radiochemical reduction.

Peptide	Variable modifications	Method
[45]GLVLIAFSQYLQQCPFDEHVK[65]	Amine-EF5 (+270.05 m/z) on **Cys58**	R, Z
[45]GLVLIAFSQYLQQCPFDEHVKLVNELTEFAK[75]	Amine-EF5 (+270.05 m/z) on **Cys58**	R
[76]TCVADESHAGCEK[88]	Amine-EF5 (+270.05 m/z) on **Cys76, 86**	R
[76]TCVADESHAGC*EK[88]	Amine-EF5 (+270.05 m/z) on **Cys76**	R, Z
[76]TC*VADESHAGCEK[88]	Amine-EF5 (+270.05 m/z) on **Cys86**	R, Z
[76]TC*VADESHAGCEKSLHTLFGDELC*K[100]	Amine-EF5 (+270.05 m/z) on **Cys86**	R
[89]SLHTLFGDELCK[100]	Amine-EF5 (+270.05 m/z) on **Cys99**	R, Z
[89]SLHTLFGDELCKVASLR[105]	Amine-EF5 (+270.05 m/z) on **Cys99**	R
[94]FGDELC(+270.05)K[100]	Amine-EF5 (+270.05 m/z) on **Cys99**	R
[123]NECFLSHKDDSPDLPK[138]	Amine-EF5 (+270.05 m/z) on **Cys125**	Z
[118]QEPERNECFLSHKDDSPDLPK[138]	Amine-EF5 (+270.05 m/z) on **Cys125**	R
[139]LKPDPNTLCDEFK[151]	Amine-EF5 (+270.05 m/z) on **Cys147**	R, Z
[139]LKPDPNTLCDEFKADEK[155]	Amine-EF5 (+270.05 m/z) on **Cys147**	R, Z
[184]YNGVFQEC*C*QAEDKGACLLPK[204]	Amine-EF5 (+270.05 m/z) on **Cys200**	R
[198]GACLLPK[204]	Amine-EF5 (+270.05 m/z) on **Cys200**	R, Z
[221]LRCASIQK[238]	Amine-EF5 (+270.05 m/z) on **Cys223**	R
[221]LRCASIQKFGER[232]	Amine-EF5 (+270.05 m/z) on **Cys223**	R
[223]CASIQKFGER[232]	Amine-EF5 (+270.05 m/z) on **Cys223**	R
[264]VHKEC*C*HGDLLECADDR[280]	Amine-EF5 (+270.05 m/z) on **Cys276**	R, Z
[264]VHKEC*C*HGDLLECADDRADLAK[285]	Amine-EF5 (+270.05 m/z) on **Cys276**	R, Z
[267]EC*C*HGDLLECADDRADLAK[285]	Amine-EF5 (+270.05 m/z) on **Cys276**	R, Z
[286]YICDNQDTISSK[297]	Amine-EF5 (+270.05 m/z) on **Cys289**	R, Z
[286]YICDNQDTISSKLK[299]	Amine-EF5 (+270.05 m/z) on **Cys289**	R
[297]LKECCDKPLLEK[309]	Amine-EF5 (+270.05 m/z) on **Cys301, 302**	R
[297]LKEC*CDKPLLEK[309]	Amine-EF5 (+270.05 m/z) on **Cys302**	Z
[297]LKEC*C*DKPLLEKSHCIAEVEK[318]	Amine-EF5 (+270.05 m/z) on **Cys312**	R
[300]ECC*DKPLLEK[309]	Amine-EF5 (+270.05 m/z) on **Cys302**	R
[302]CDKPLLEK[309]	Amine-EF5 (+270.05 m/z) on **Cys302**	R
[310]SHCIAEVE[317]	Amine-EF5 (+270.05 m/z) on **Cys312**	R
[310]SHCIAEVEK[318]	Amine-EF5 (+270.05 m/z) on **Cys312**	R, Z
[310]SHCIAEVEKD[319]	Amine-EF5 (+270.05 m/z) on **Cys312**	R
[310]SHCIAEVEKDAIPENLPPLTADFAEDK[336]	Amine-EF5 (+270.05 m/z) on **Cys312**	R

Peptide	Variable modifications	Method
^{310}SHCIAEVEKDAIPENLPPLTADFAEDKDVC*^{339}	Amine-EF5 (+270.05 m/z) on **Cys312**	R, Z
^{310}SHC*IAEVEKDAIPENLPPLTADFAEDKDVCK^{340}	Amine-EF5 (+270.05 m/z) on **Cys339**	R, Z
^{319}DAIPENLPPLTADFAEDKDVCK^{340}	Amine-EF5 (+270.05 m/z) on **Cys339**	R, Z
^{319}DAIPENLPPLTADFAEDKDVCKNYQEAK^{347}	Amine-EF5 (+270.05 m/z) on **Cys339**	R
^{387}DDPHACYSTVFDKLK^{401}	Amine-EF5 (+270.05 m/z) on **Cys392**	R, Z
^{402}HLVDEPQNLIKQNCDQFEK^{420}	Amine-EF5 (+270.05 m/z) on **Cys415**	R
^{402}HLVDEPQNLIKQNCDQFEKLGEYGFQNALIVR^{433}	Amine-EF5 (+270.05 m/z) on **Cys415**	R
^{413}QNCDQFEK^{420}	Amine-EF5 (+270.05 m/z) on **Cys415**	R, Z
^{413}QNCDQFEKLGEYGFQNALIVR^{433}	Amine-EF5 (+270.05 m/z) on **Cys415**	R, Z
^{456}VGTRCCTKPESER^{468}	Amine-EF5 (+270.05 m/z) on **Cys460, 461**	R
^{460}C*C*TKPESERM#PCTEDYLSLILNR^{482}	Amine-EF5 (+270.05 m/z) on **Cys471**	R
^{469}M#PCTEDYLSLILNR^{482}	Amine-EF5 (+270.05 m/z) on **Cys471**	R, Z
^{483}LCVLHEK^{489}	Amine-EF5 (+270.05 m/z) on **Cys471**	R, Z
^{483}LCVLHEKTPVSEK^{495}	Amine-EF5 (+270.05 m/z) on **Cys471**	R
^{483}LCVLHEKTPVSEKVTK^{498}	Amine-EF5 (+270.05 m/z) on **Cys471**	R
^{489}TPVSEKVTKCCTESLVNR^{507}	Amine-EF5 (+270.05 m/z) on **Cys499, 500**	R
^{496}VTKC*CTESLVNR^{507}	Amine-EF5 (+270.05 m/z) on **Cys500**	R
^{496}VTKC*CTESLVNRRPC*^{510}	Amine-EF5 (+270.05 m/z) on **Cys500**	R
^{499}C*CTESLVNRRPCFSALTPDETYVPK^{523}	Amine-EF5 (+270.05 m/z) on **Cys500, 510**	R
^{507}RRPCFSALTPDETYVPK^{523}	Amine-EF5 (+270.05 m/z) on **Cys510**	R
^{508}RPCFSALTPDETYVPK^{523}	Amine-EF5 (+270.05 m/z) on **Cys510**	R, Z
^{508}RPCFSALTPDETYVPKAFDEK^{528}	Amine-EF5 (+270.05 m/z) on **Cys510**	R
^{509}PCFSALTPDETYVPK^{523}	Amine-EF5 (+270.05 m/z) on **Cys471**	R
^{524}AFDEKLFTFHADICTLPDTEK^{544}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{524}AFDEKLFTFHADICTLPDTEKQIK^{547}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{529}LFTFHADICTLPDTEK^{544}	Amine-EF5 (+270.05 m/z) on **Cys537**	R, Z
^{529}LFTFHADICTLPDTEKQIK^{547}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{533}HADICTLPDTEK^{544}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{534}ADICTLPDTEK^{544}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{535}DICTLPDTEK^{544}	Amine-EF5 (+270.05 m/z) on **Cys537**	R
^{581}C*C*AADDKEACFAVEGPK^{597}	Amine-EF5 (+270.05 m/z) on **Cys590**	R, Z
^{588}EACFAVEGPK^{597}	Amine-EF5 (+270.05 m/z) on **Cys590**	R, Z

Supplemental table 5. Mass spectrometric identification of nitroimidazole adducts on lysozyme.
List of modified peptides that were found for the trypsin digestion of lysozyme adducts. # denotes a oxidation modification (+16.00 Da) and * denotes a carbamidomethylation (+57.02 Da). Z indicates zinc reduction; R indicates radiochemical reduction.

Peptide	Variable modifications	Method
^{24}CELAAAMK^{31}	Amine-EF5 (+270.05 m/z) on **Cys24** Amine-Pimo (+225.15 m/z) on **Cys24**	R R
^{24}CELAAAM#K^{31}	Amine-EF5 (+270.05 m/z) on **Cys24** Amine-Pimo (+225.15 m/z) on **Cys24**	R, Z R
^{24}CELAAAMKR^{32}	Amine-EF5 (+270.05 m/z) on **Cys24** Amine-Pimo (+225.15 m/z) on **Cys24**	R R
^{24}CELAAAM#KR^{32}	Amine-EF5 (+270.05 m/z) on **Cys24** Amine-Pimo (+225.15 m/z) on **Cys24**	R, Z R
^{40}GYSLGNWVCAAK^{51}	Amine-EF5 (+270.05 m/z) on **Cys48** Amine-Pimo (+225.15 m/z) on **Cys48**	R, Z R
^{80}WWCNDGR^{86}	Amine-EF5 (+270.05 m/z) on **Cys82**	R, Z
^{80}WWCNDGRTPGSR^{91}	Amine-EF5 (+270.05 m/z) on **Cys82**	R, Z
^{92}NLCNIPC*SALLSSDITASVNC*AK^{114}	Amine-EF5 (+270.05 m/z) on **Cys94** Amine-Pimo (+225.15 m/z) on **Cys94**	R, Z Z
^{92}NLC*NIPCSALLSSDITASVNC*AK^{114}	Amine-EF5 (+270.05 m/z) on **Cys98** Amine-Pimo (+225.15 m/z) on **Cys98**	R, Z Z
^{92}NLC*NIPC*SALLSSDITASVNCAK^{114}	Amine-EF5 (+270.05 m/z) on **Cys112** Amine-Pimo (+225.15 m/z) on **Cys112**	R, Z R, Z
^{92}NLCNIPC*SALLSSDITASVNC*AKK^{115}	Amine-Pimo (+225.15 m/z) on **Cys94**	R
^{92}NLC*NIPCSALLSSDITASVNC*AKK^{115}	Amine-EF5 (+270.05 m/z) on **Cys98** Amine-Pimo (+225.15 m/z) on **Cys98**	Z Z
^{92}NLC*NIPC*SALLSSDITASVNCAKK^{115}	Amine-EF5 (+270.05 m/z) on **Cys112**	Z
^{131}NRCKGTDVQAWIR^{143}	Amine-EF5 (+270.05 m/z) on **Cys133**	Z
^{133}CKGTDVQAWIR^{143}	Amine-EF5 (+270.05 m/z) on **Cys133** Amine-Pimo (+225.15 m/z) on **Cys133**	Z R

Supplemental references

1. Lord, E.M., L. Harwell, and C.J. Koch, Detection of hypoxic cells by monoclonal antibody recognizing 2-nitroimidazole adducts. Cancer Res, 1993. 53(23): p. 5721-6.
2. Koch, C.J. and J.A. Raleigh, Radiolytic reduction of protein and nonprotein disulfides in the presence of formate: a chain reaction. Arch Biochem Biophys, 1991. 287(1): p. 75-84.
3. Raleigh, J.A. and C.J. Koch, Importance of thiols in the reductive binding of 2-nitroimidazoles to macromolecules. Biochem Pharmacol, 1990. 40(11): p. 2457-64.
4. Varghese, A.J. and G.F. Whitmore, Binding of nitroreduction products of misonidazole to nucleic acids and protein. Cancer Clin Trials, 1980. 3(1): p. 43-6.
5. Masaki, Y., et al., FMISO accumulation in tumor is dependent on glutathione conjugation capacity in addition to hypoxic state. Ann Nucl Med, 2017. 31(8): p. 596-604.
6. Varghese, A.J., Glutathione conjugates of misonidazole. Biochem Biophys Res Commun, 1983. 112(3): p. 1013-20.
7. Kizaka-Kondoh, S. and H. Konse-Nagasawa, Significance of nitroimidazole compounds and hypoxia-inducible factor-1 for imaging tumor hypoxia. Cancer Sci, 2009. 100(8): p. 1366-73.
8. Fleming, I.N., et al., Imaging tumour hypoxia with positron emission tomography. Br J Cancer, 2015. 112(2): p. 238-50.

CHAPTER 7

A COMPARISON OF MACHINE LEARNING MODELS VERSUS CLINICAL EVALUATION FOR MORTALITY PREDICTION IN PATIENTS WITH SEPSIS

William P.T.M. van Doorn, Patricia M. Stassen, Hella F. Borggreve, Maaike J. Schalkwijk, Judith Stoffers, Otto Bekers, Steven J.R. Meex

PLOS ONE
2021 Jan 19;16(1):e0245157

Abstract

Introduction: Patients with sepsis who present to an emergency department (ED) have highly variable underlying disease severity, and can be categorized from low to high risk. Development of a risk stratification tool for these patients is important for appropriate triage and early treatment. The aim of this study was to develop machine learning models predicting 31-day mortality in patients presenting to the ED with sepsis and to compare these to internal medicine physicians and clinical risk scores.

Methods: A single-center, retrospective cohort study was conducted amongst 1,344 emergency department patients fulfilling sepsis criteria. Laboratory and clinical data that was available in the first two hours of presentation from these patients were randomly partitioned into a development (n=1,244) and validation dataset (n=100). Machine learning models were trained and evaluated on the development dataset and compared to internal medicine physicians and risk scores in the independent validation dataset. The primary outcome was 31-day mortality.

Results: A number of 1,344 patients were included of whom 174 (13.0%) died. Machine learning models trained with laboratory or a combination of laboratory + clinical data achieved an area-under-the ROC curve of 0.82 (95% CI: 0.80-0.84) and 0.84 (95% CI: 0.81-0.87) for predicting 31-day mortality, respectively. In the validation set, models outperformed internal medicine physicians and clinical risk scores in sensitivity (92% vs. 72% vs. 78%;$p<0.001$,all comparisons) while retaining comparable specificity (78% vs. 74% vs. 72%;$p>0.02$). The model had higher diagnostic accuracy with an area-under-the-ROC curve of 0.85 (95%CI: 0.78-0.92) compared to abbMEDS (0.63,0.54-0.73), mREMS (0.63,0.54-0.72) and internal medicine physicians (0.74,0.65-0.82).

Conclusion: Machine learning models outperformed internal medicine physicians and clinical risk scores in predicting 31-day mortality. These models are a promising tool to aid in risk stratification of patients presenting to the ED with sepsis.

Introduction

Among emergency department (ED) presentations, a substantial number of patients present with symptoms of sepsis [1]. Sepsis is defined as a systemic inflammatory response syndrome (SIRS) to an infection and is associated with a wide variety of risks including septic shock and death [2]. Mortality rates of sepsis are as high as 16%, potentially increasing up to 40% when suffering from septic shock [2, 3]. Novel clinical decision support (CDS) systems capable of identifying low- or high-risk patients could become important for early treatment and triage of ED patients, but also for preventing unnecessary referrals to the intensive care unit (ICU). EDs are one of the most overcrowded units of a modern hospital, highlighting the importance of proper allocation and management of resources [1]. Development of a risk stratification tool for patients with sepsis may improve health outcome in this group, but may also contribute to resolve the problem of overcrowded EDs.

Currently, a wide variety of clinical risk scores are used in routine clinical care to facilitate risk stratification of patients with sepsis [4]. These include the relatively simple (quick) sequential organ failure assessment ((q)SOFA) score [5, 6], but also more complex scores such as the abbreviated Mortality in Emergency Department Sepsis (abbMEDS) score and modified Rapid Emergency Medicine Score (mREMS) [7, 8]. These traditional risk scores have shown varying performance for predicting 28-day mortality (area under the receiver operating characteristic curve (AUC) for abbMEDS: 0.62-0.85, mREMS: 0.62-0.84 and SOFA: 0.61-0.82) [3, 8-11]. In addition, clinical judgment of the attending physician in the ED plays an important role in risk stratification. The judgment of physicians was found to be a moderate to good predictor (AUC of 0.68-0.81) of mortality in the ED [12, 13].

Interestingly, a new group of CDS systems are being developed based on machine learning (ML) technology [14]. Machine learning can extract information from complex, non-linear data and provide insights to support clinical decision making. Hence, the first studies emerged that report machine learning-based mortality prediction models using data from patients with sepsis presenting to the ED [15-26]. Unfortunately, these studies did not provide a comparison with physicians in terms of prognostic performance. Recently, a new group of machine learning algorithms termed gradient boosting trees emerged; showing superior performance compared to other ML models in some problems within the medical domain [27, 28]. Exploring if these models can outperform clinical risk scores and clinical judgment of physicians in their ability to identify low- or high-risk patients is a necessary step to explore the potential value of machine learning models in clinical practice.

The aim of this study was to develop machine learning-based prediction models for all-cause mortality at 31 days based on available laboratory and clinical data from patients presenting to the ED with sepsis. Subsequently, we compared the performance of these machine learning models with judgment of internal medicine physicians and clinical risk scores; abbMEDS, mREMS and SOFA.

Methods

Study design and setting

We performed a retrospective cohort study among all patients who presented to the ED at the Maastricht University Medical Centre+ between January 1, 2015 and December 31, 2016. All patients aged ≥18 years being referred to the internal medicine physician with sepsis, defined as a proven or suspected infection, and two or more SIRS and/or qSOFA criteria (S1 supporting information) were included in this study [2, 5, 29]. Patients with missing clinical data or with less than four laboratory results were excluded. Also, patients who refused to give consent were excluded. This study was approved by the medical ethical committee (METC 2019-1044) and the hospital board of the Maastricht University Medical Centre+. Furthermore, the study follows the STROBE guidelines and was conducted according to the principles of the Declaration of Helsinki [30]. The ethics committee waived the requirement for informed consent.

Data collection and processing

We collected clinical and laboratory data from all patients included in the study available within two hours after initial ED presentation. Clinical data were manually extracted through the electronic health record of the patient and included characteristics such as vital signs, hemodynamic parameters, and medical history (S1 Table). Biomarkers requested for standard clinical care were acquired through the laboratory information system. Biomarkers that were ordered in less than 1/1000 patients were excluded from the analysis. A list of included biomarkers is provided in S1 Table. Missing values did not require any processing as our machine learning model is capable of dealing with missing data. Instead, we created an additional variable for each biomarker with a discrete 'absence' or 'presence' feature to enable our model to distinguish between the absence and presence of a laboratory test within a patient. These features were included in both datasets. Finally, we derived two datasets from the processed data:

Laboratory dataset: this dataset consisted of age, sex, time of laboratory request and all requested laboratory biomarkers within two hours after the initial laboratory request

Laboratory + clinical dataset: this dataset contained all variables from the laboratory dataset, and additionally clinical, vital and physical (e.g. length and weight) characteristics of the patient

A full overview of all variables present in each dataset is described in S1 Table. Datasets were anonymized and randomly divided into two subsets: 1) a development subset (n=1,244), used for model training and evaluation, and 2) an independent validation subset (n=100), used for final validation and comparison of models with judgment of acute internal medicine physicians and clinical risk scores; abbMEDS and mREMS. A schematic overview of the study design and model development is depicted in Fig 1. Data processing and manipulation was performed using Python programming language (version 3.7.1) using packages numpy (version 1.17) and Pandas (version 0.24).

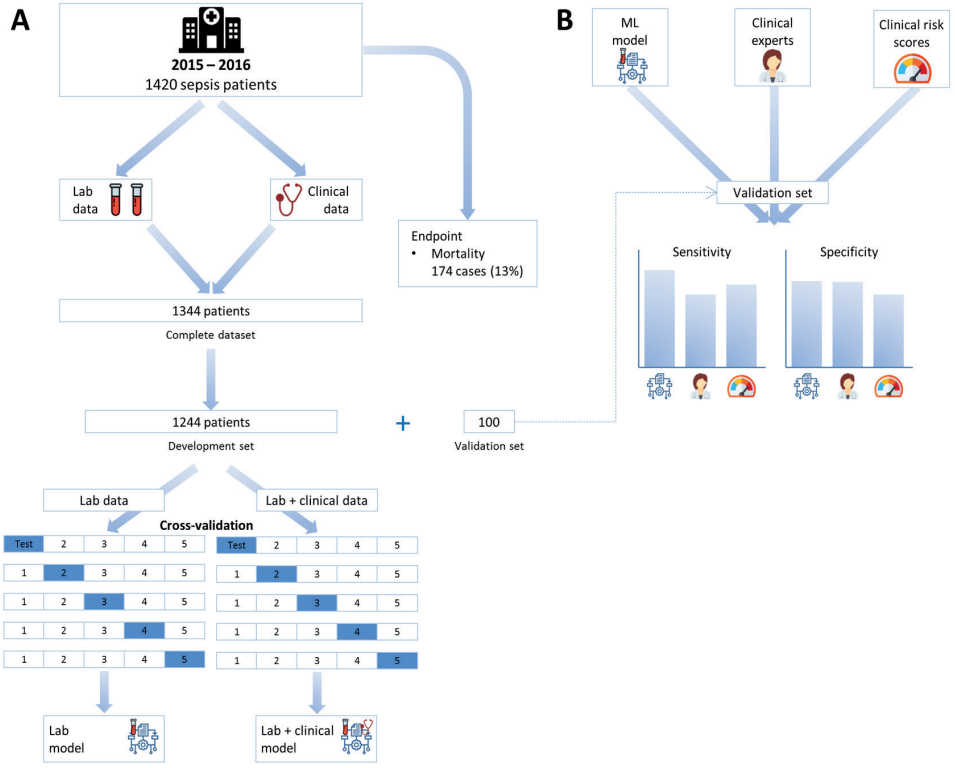


Figure 1. Overview of study design and model development. (A) We included 1,344 patients with a diagnosis of sepsis who presented to the ED. Patients were randomly partitioned in a development subset (n=1,244), used to train and evaluate performance of machine learning models, and a validation subset (n=100), used to compare models with internal medicine physicians and clinical risk scores. Cross-validation was used to obtain a robust estimate of model performance in the development subset. (B) The machine learning model with the highest cross-validation performance was compared internal medicine physicians and clinical risk scores to predict 31-days mortality.

Outcome measure

Septic shock during presentation was defined as systolic blood pressure (SBP) ≤90 mmHg and mean arterial pressure (MAP) ≤65 mmHg despite adequate fluid resuscitation. The outcome measure for this study was death within 31 days (1 month) after initial ED presentation. All-cause mortality information was acquired through electronic health records.

Model training and evaluation

Our proposed predictive model uses individual patient data available within two hours after initial ED presentation and generates the probability of mortality within 31 days. This prediction task can be solved by a variety of statistical and machine learning models. In the current study we evaluated logistic regression, random forest, multi-layer perceptron neural networks and XGBoost (S2 supporting information and S2 Table) on the laboratory dataset. We selected XGBoost as our machine learning model of choice as this was proven to possess the highest baseline performance (S2 Table). XGBoost is a recent implementation of gradient tree boosting systems which involve combining the predictions of many "weak" decision trees into a strong predictor [27]. This recent implementation is characterized by integral support of missing data and regularization mechanisms to prevent overfitting [27]. XGBoost models and their development can be altered by adjusting the parameters of the technique, referred to as "hyperparameters". Due to sample size limitations and the scope of our study, we decided not to optimize our hyperparameters and predefined them as described in S3 Table.

We employed stratified K-fold cross validation to assess the generalizability of our prediction models. Briefly, we randomly partitioned the development subset (n=1,244) into five, equally sized, folds. During each round of cross-validation, four of these folds were used to train our models ("train set") and the fifth was used to evaluate performance ("test set"). This was done in such a manner that every fold would be labeled as test set only once. We monitored training and test set errors to ensure that training increased performance on the test set. Accordingly, training was terminated after 5,000 rounds or when performance on the test set did not further improve for 10 rounds. We evaluated developed models trained with (i) the laboratory dataset or (ii) the laboratory + clinical dataset, resulting in a total of two independent cross-validations.

Model explanation

To explain the output of our XGBoost models, we used the SHapley Additive exPlanations (SHAP) algorithm, to help us understand how a single feature affects the output of the model [31-33]. SHAP uses a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions [34, 35]. A Shapley

value states, given the current set of variables, how much a variable in the context of its interaction with other variables contributes to the difference between the actual prediction and the mean prediction. That is, the mean prediction plus the sum of the Shapley values for all variables equals the actual prediction. It is important to understand that this is fundamentally different to direct variable effects known from e.g. (generalized) linear models. The SHAP value for a variable should not be seen as its direct -and isolated effect- but as its aggregated effect when interacting with other variables in the model. In our specific case, positive Shapley values contribute towards a positive prediction (death), whilst low or negative Shapely values contribute towards a negative prediction (survival). ML training and evaluation was done in Python using packages Keras (version 2.2.2), XGBoost (version 0.90), SHAP (version 0.34.0) and scikit-learn (version 0.22.1). The analysis code for this study is available on reasonable request.

Comparison of machine learning with internal medicine physicians and clinical risk scores

Performance of machine learning models was compared with clinical judgment of acute internal medicine physicians (n=4) and clinical risk scores in a validation subset of patients with sepsis (n=100) which were not previously exposed to the ML model. We selected the best performing machine learning model from cross-validation and trained this with identical hyperparameters as previously described on the full development subset. A machine learning prediction of higher than 0.50 was considered as a positive prediction. Next, we calculated the mREMS, abbMEDS and SOFA clinical risk scores as described previously (S1 supporting information) [8, 36]. Acute internal medicine physicians (n=4; 2 experienced consultants in acute internal medicine and 2 experienced residents acute internal medicine) were asked to predict 31-day mortality in the validation subset, based on retrospectively collected clinical and laboratory data. This data was presented in the form of a simulated electronic health record.

Statistical analysis

Descriptive analysis of baseline characteristics was performed using IBM SPSS Statistics for Windows (version 24.0). Continuous variables were reported as means with standard deviation (SD) or medians with interquartile ranges (IQRs) depending on the distribution of the data. Categorical variables were reported as proportions. Cross-validated models were assessed by receiver operating characteristic (ROC) curves and compared by their AUC using the Wilcoxon matched-pairs signed rank test. Besides diagnostic performance, we assessed calibration in cross-validations with reliability curves [37] and brier scores [38]. In our final validation subset, we compared the predictive performance of our best performing ML model to the judgment of acute internal medicine physicians and clinical risk scores with respect to sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and AUC. Differences in AUC were tested using the method of DeLong et al [39]. Confidence intervals for proportions (e.g. sensitivity) were calculated

using binomial testing and compared using McNemar's test. To analyze individual differences between internal medicine physicians, we performed two additional sensitivity analyses. First, the Cohen κ statistic was used to measure the inter-observer agreement between the internal medicine physicians. The level of agreement was interpreted as nil if κ was 0 to 0.20; minimal, 0.21 to 0.39; weak, 0.40 to 0.59; moderate, 0.60 to 0.79; strong, 0.80 to 0.90; and almost perfect, 0.90 to 1 [40]. Second, we compared the machine learning model against alternating groups of internal medicine physicians in which one physician was removed in each comparison.

Results

Study population and characteristics

During the study period, 5,967 patients presented to the ED who were referred to an internal medicine physician in our hospital. Of these patients, we included 1,420 patients with a suspected or proven infection, fulfilling the SIRS and/or qSOFA criteria. A number of 76 patients were excluded due to missing clinical data (n=23) and insufficient number of laboratory results (n=53), to form a final cohort of 1,344 patients (S1 Fig). Among all patients, 102 (7.6%) suffered from septic shock during presentation at ED and 174 (13.0%) died within 31 days after initial ED presentation. Baseline characteristics of the study patients in development and validation datasets are shown in Table 1.

Table 1. Baseline characteristics of patients in the development and validation datasets.

Characteristics	Development N = 1.244	Validation N = 100
Demographics		
Age	71.3 (58.8-82.3)	70.8 (58.4-82.8)
Sex, female	567 (45.6)	58 (58.0)
Comorbidity		
Cancer	446 (35.9)	28 (28.0)
Cardiopulmonary	381 (30.6)	30 (30.0)
Diabetes	264 (21.2)	19 (19.0)
Renal disease	128 (10.3)	9 (9.0)
Liver disease	42 (3.4)	7 (7.0)
Neuropsychiatric	65 (5.2)	2 (2.0)
Focus of infection at ED		
Respiratory tract	421 (33.8)	34 (34.0)
Urinary tract	218 (17.5)	18 (18.0)
Gastrointestinal tract	415 (33.4)	37 (37.0)
Others	75 (6.0)	6 (6.0)
Skin	115 (9.2)	5 (5.0)
Severity scores		
abbMEDS[a]	5.5 (3-8)	6 (3-8)
mREMS[b]	7 (6-9)	7 (6-9)
SOFA[c]	7 (5-9)	6 (5-8)
Outcomes		
Septic shock	94 (7.6)	8 (8.0)
31-day mortality	161 (12.9)	13 (13.0)

[a] *AbbMEDS, Abbreviated Mortality in ED Sepsis, was calculated as described by Vorwerk et al [8].*
[b] *mREMS, modified Rapid Emergency Medicine Score, was calculated as described by Chang et al [36].*
[c] *SOFA, Sepsis-related Organ Failure Assessment, was calculated as described by Vincent et al [6].*

Machine learning development and evaluation

To assess the generalizability of our developed XGBoost models, we employed five-fold cross validation on the development dataset (n=1,244). XGBoost models trained with laboratory data achieved an AUC of 0.82 (95% CI: 0.80 – 0.84) for predicting 31-day mortality (Fig 2). The performance improved, although not statistically significant, when clinical data was added to the laboratory data to train XGBoost to an AUC of 0.84 (95% CI: 0.81 – 0.87) for mortality (compared to lab only; p=0.25). Individual cross-validation results of each model are depicted in S2 Fig. Additionally, calibration curves show well calibrated models with brier scores between 0.08 to 0.10 (S3 Fig).

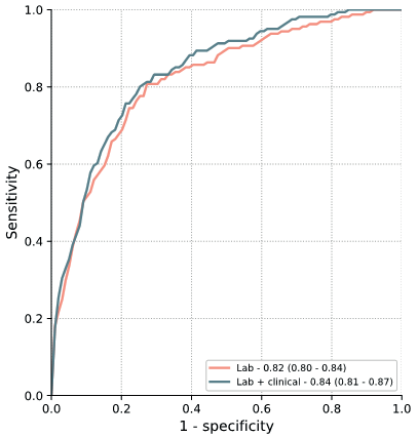


Figure 2. XGBoost model performance for predicting all-cause mortality at 31 days in the development dataset. Models trained with laboratory data achieved a mean AUC of 0.82 (95% CI: 0.80 – 0.84) for predicting 31-day mortality. Predictive performance increased when models were trained with laboratory + clinical data to a mean AUC of 0.84 (95% CI: 0.81 – 0.87), but this was not statistically different (p=0.25).

Model explanation

To identify which laboratory and clinical features contributed most to the performance of our models, we calculated SHAP values for the (i) laboratory and (ii) laboratory + clinical models (Fig 3). Among the highest ranked features, we observe features that are also often used in risk scores including urea, platelet count, glasgow coma score (GCS) and blood pressure. Interestingly, we also observe features such as glucose, lipase, and GCS which are less commonly associated with mortality in sepsis patients. An extended analysis of the correlation between important features in our models and risk scores is provided in S4 Table. Moreover, these SHAP plots allow us to examine the individual impact of laboratory and clinical features on the predictions of our models. For example, higher urea and C-reactive protein (CRP) levels (represented by red points) have a high SHAP value and thus a positive effect on the model outcome (death).

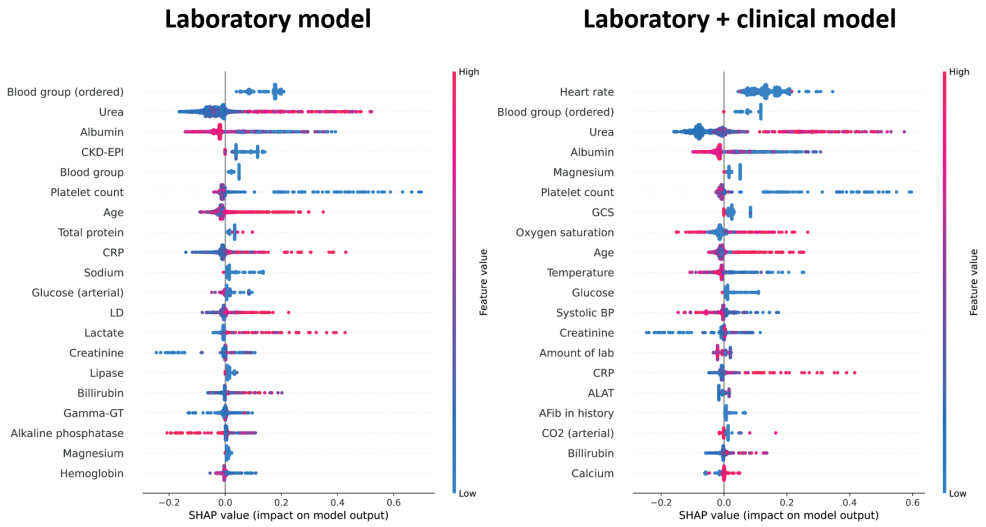


Figure 3. Analysis of parameter importance in the XGBoost models. Models with laboratory data (left) and with laboratory + clinical data (right) were analyzed using SHAP values. Individual parameters are ranked by importance in descending order based on the sum of the SHAP values over all the samples. Negative or low SHAP values contribute towards a negative model outcome (survival), whereas high SHAP values contribute towards a positive model outcome (death).

Machine learning versus internal medicine physicians and clinical risk scores

To explore the potential value of machine learning models in clinical practice, we compared the model trained with laboratory + clinical data with acute internal medicine physicians and clinical risk scores, abbMEDS, mREMS and SOFA, to predict 31-day mortality. In an independent validation subset (n=100) -which the model never had been exposed to before- it achieved a sensitivity of 0.92 (95% CI: 0.87-0.95, Fig 4A) and specificity of 0.78 (95% CI: 0.70-0.86, Fig 4B). In terms of sensitivity, the machine learning model significantly outperformed internal medicine physicians (0.72, 95% CI: 0.62-0.81; p<0.001), abbMEDS (0.54, 95% CI: 0.44-0.64; p<0.0001), mREMS (0.62, 95% CI: 0.52-0.72; p<0.001) and SOFA (0.77, 95% CI: 0.69-0.85; p=0.003). On the other hand, the model retained a specificity that was comparable to that of internal medicine physicians (0.74, 95% CI: 0.64-0.82; p=0.509), abbMEDS (0.72, 95% CI: 0.64-0.81; p=0.327) and SOFA (0.74, 95% CI: 0.65-0.82, p=0.447), while still outperforming mREMS (0.64, 95% CI: 0.55-0.74; p=0.02). Additionally, the model had higher overall diagnostic accuracy with an AUC of 0.852 (95% CI: 0.783-0.922) compared to abbMEDS (0.631, 0.537-0.726, p=0.021), mREMS (0.630, 0.535-0.724, p=0.016), SOFA (0.752, 0.667-0.836, p=0.042) and internal medicine physicians (0.735, 0.648-0.821, p=0.032-0.189) (S4 Fig and S5 Table). Similar observations were made in additional evaluation metrics such as positive predictive value (NPV), negative predictive value (NPV) and accuracy (S5 Table). Individually, consultants were found to be more

sensitive compared to residents (S5 Fig) with a poor to moderate agreement between the internists (Cohen's Kappa 0.46 to 0.67) (S6 Table). A sensitivity analysis with four additional comparisons, where one physician was excluded at a time, confirmed that the results are robust and that the outperformance of the machine learning model was not due to an outlier in the physician group (S7 Table).

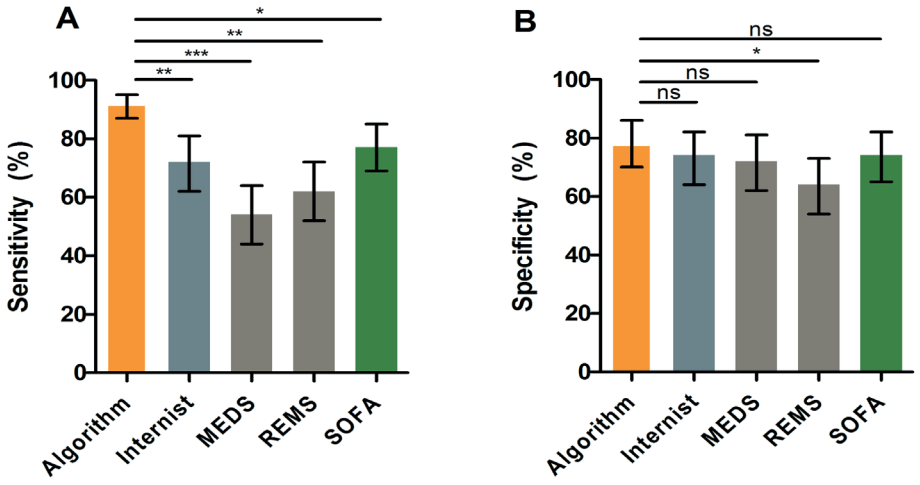


Figure 4. Comparison of XGBoost model with internal medicine physicians and clinical risk scores. The XGBoost model achieved a sensitivity (A) of 0.92 (95% CI: 0.87-0.95) and specificity (B) of 0.78 (95% CI: 0.70-0.86) for predicting mortality. This was significantly better than the mean prediction of internal medicine physicians for sensitivity (0.72, 0.62-0.81; p<0.001) as well as abbMEDS (0.54, 0.44-0.64; p<0.0001), mREMS (0.62, 0.52-0.72; p<0.001) and SOFA (0.77, 95% CI: 0.69-0.85; p=0.003). In terms of specificity, internal medicine physicians (0.74, 0.64-0.82; p=0.509), abbMEDS (0.72, 0.64-0.81; p=0.327) and SOFA (0.74, 95% CI: 0.65-0.82, p=0.447) achieved similar performance compared to the XGBoost model, opposed to mREMS (0.64, 0.55-0.74; p=0.02) which was significantly worse than machine learning predictions. * = p<0.05; ** = p<0.001; *** = p<0.0001; NS = not significant.

Discussion

In the present study we demonstrate the application of machine learning models to predict 31-day mortality patients presenting to the ED with sepsis. Our study reports several important findings.

First, we show that machine learning based models can accurately predict 31-day mortality in patients with sepsis. Highest diagnostic accuracy was obtained with the model that was trained with both laboratory and clinical data. Patient characteristics that are employed in traditional risk scores, such as blood pressure and heart rate, were also found to be amongst the most important variables for model predictions. Second, machine learning models outperformed the judgment of internal medicine physicians and commonly used clinical risk scores, abbMEDS, mREMS and SOFA. Specifically, machine learning was more sensitive compared with risk scores and internal medicine physicians, while retaining identical or slightly higher specificity. These preliminary data provide support in favor of the development and implementation of machine learning based models as clinical decision support tools, e.g. risk stratification of sepsis patients presenting to the ED.

We are aware of several studies which describe the machine-learning based prediction of mortality in sepsis populations presenting to the ED [15-17]. Taylor et al. described a random forest model outperforming clinical risk scores in an ED population. Despite their bigger population, our XGBoost model appears to achieve similar performance to their random forest model, which corroborates and extends the power of this machine learning technique. Two recent studies by Barnaby et al. and Chiew et al. focused on using heart rate variability (HRV) for risk prediction in sepsis patients and reported predictive performance similar to our findings [15, 16]. Interestingly, their populations were smaller and this would therefore also advocate the use of HRV in our models. Despite these findings, Chiew et al. demonstrated that models without laboratory data significantly decreased in performance, emphasizing the importance of laboratory data in these machine learning models. Nevertheless, to the best of our knowledge this is the first study to report the direct comparison of machine learning models with internal medicine physicians. Although we do not present prospective results, we demonstrate that machine learning outperforms clinical judgment of internal medicine physicians and clinical risk scores, implying that current XGBoost models potentially aid in risk stratification of ED patients. As an example, implementation of these models should revolve around identifying patients with a high risk, e.g. ≥50% mortality within 31 days, which would then be re-evaluated once more before being discharged from the ED. This kind of implementation was shown in a recent randomized clinical trial by Shimabukuro et al. [41], proving that average length of stay and in-hospital mortality decreased by using a ML-based sepsis detection model in the ICU. Although this was carried out with a small population in an ICU instead of the ED, it clearly shows the potential of ML-based risk stratifying models.

The current study has several strengths and limitations. Strengths include (i) comparison of laboratory versus laboratory + clinical models, (ii) analysis of features contributing to models' prediction and (iii) the comparison with internal medicine specialists. We are also aware of several limitations. First, the present study was a single-center study with a relatively small sample size at least from a machine learning analysis perspective. Nearly all machine learning models scale exceptionally well with data, and therefore substantial further improvement of diagnostic accuracy is likely when increasing the sample size. We also limited ourselves to sepsis patients presenting to the ED, and thus it is unknown to what degree these models translate to a broader, general ED population. Second, results presented in this study are based on retrospective data in a single center, limiting the external validity of the model. Unfortunately, this limitation currently applies to most studies applying ML in medicine. Third, the present study focused on model development and subsequent performance comparison with clinical judgment and clinical risk scores. It should be noted that the comparison with internal medicine specialists was performed using retrospectively generated electronic health records, rather than a prospective evaluation, which might have underestimated their diagnostic performance as they were not able to directly "see" the patient. Prospective evaluation, in respect to mortality, but also in relation to clinical endpoints that confirm true clinical benefit would facilitate implementation of ML-based risk stratification tools in clinical practice.

Conclusion

In conclusion, the present proof-of-concept study demonstrates the potential of machine learning models to predict mortality in patients with sepsis presenting to the ED. Machine learning outperformed clinical judgment of internal medicine physicians and established clinical risk scores. These data provide support in favor of the implementation of machine learning based risk stratification tools of sepsis patients presenting to the ED.

References

1. LaCalle E, Rabin E. Frequent users of emergency departments: the myths, the data, and the policy implications. Ann Emerg Med. 2010;56(1):42-8. Epub 2010/03/30. doi: 10.1016/j.annemergmed.2010.01.032. PubMed PMID: 20346540.
2. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10. Epub 2016/02/24. doi: 10.1001/jama.2016.0287. PubMed PMID: 26903338; PubMed Central PMCID: PMCPMC4968574.
3. Roest AA, Tegtmeier J, Heyligen JJ, Duijst J, Peeters A, Borggreve HF, et al. Risk stratification by abbMEDS and CURB-65 in relation to treatment and clinical disposition of the septic patient at the emergency department: a cohort study. BMC Emerg Med. 2015;15:29. Epub 2015/10/16. doi: 10.1186/s12873-015-0056-z. PubMed PMID: 26464225; PubMed Central PMCID: PMCPMC4605126.
4. McLymont N, Glover GW. Scoring systems for the characterization of sepsis and associated outcomes. Ann Transl Med. 2016;4(24):527. Epub 2017/02/06. doi: 10.21037/atm.2016.12.53. PubMed PMID: 28149888; PubMed Central PMCID: PMCPMC5233540.
5. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762-74. Epub 2016/02/24. doi: 10.1001/jama.2016.0288. PubMed PMID: 26903335; PubMed Central PMCID: PMCPMC5433435.
6. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonça A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Medicine. 1996;22(7):707-10. doi: 10.1007/BF01709751.
7. Olsson T, Terent A, Lind L. Rapid Emergency Medicine score: a new prognostic tool for in-hospital mortality in nonsurgical emergency department patients. J Intern Med. 2004;255(5):579-87. Epub 2004/04/14. doi: 10.1111/j.1365-2796.2004.01321.x. PubMed PMID: 15078500.
8. Vorwerk C, Loryman B, Coats TJ, Stephenson JA, Gray LD, Reddy G, et al. Prediction of mortality in adult emergency department patients with sepsis. Emerg Med J. 2009;26(4):254-8. Epub 2009/03/25. doi: 10.1136/emj.2007.053298. PubMed PMID: 19307384.
9. Crowe CA, Kulstad EB, Mistry CD, Kulstad CE. Comparison of severity of illness scoring systems in the prediction of hospital mortality in severe sepsis and septic shock. J Emerg Trauma Shock. 2010;3(4):342-7. Epub 2010/11/11. doi: 10.4103/0974-2700.70761. PubMed PMID: 21063556; PubMed Central PMCID: PMCPMC2966566.
10. Olsson T, Terent A, Lind L. Rapid Emergency Medicine Score can predict long-term mortality in nonsurgical emergency department patients. Acad Emerg Med. 2004;11(10):1008-13. Epub 2004/10/07. doi: 10.1197/j.aem.2004.05.027. PubMed PMID: 15466141.
11. Minne L, Abu-Hanna A, de Jonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit Care. 2008;12(6):R161. Epub 2008/12/19. doi: 10.1186/cc7160. PubMed PMID: 19091120; PubMed Central PMCID: PMCPMC2646326.
12. Rohacek M, Nickel CH, Dietrich M, Bingisser R. Clinical intuition ratings are associated with morbidity and hospitalisation. Int J Clin Pract. 2015;69(6):710-7. Epub 2015/02/18. doi: 10.1111/ijcp.12606. PubMed PMID: 25689155; PubMed Central PMCID: PMCPMC5024066.
13. Zelis N, Mauritz AN, Kuijpers LIJ, Buijs J, de Leeuw PW, Stassen PM. Short-term mortality in older medical emergency patients can be predicted using clinical intuition: A prospective study. PLoS One. 2019;14(1):e0208741. Epub 2019/01/03. doi: 10.1371/journal.pone.0208741. PubMed PMID: 30601815; PubMed Central PMCID: PMCPMC6314634.
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56. Epub 2019/01/09. doi: 10.1038/s41591-018-0300-7. PubMed PMID: 30617339.

15. Barnaby DP, Fernando SM, Herry CL, Scales NB, Gallagher EJ, Seely AJE. Heart Rate Variability, Clinical and Laboratory Measures to Predict Future Deterioration in Patients Presenting With Sepsis. Shock. 2019;51(4):416-22. Epub 2018/05/31. doi: 10.1097/SHK.0000000000001192. PubMed PMID: 29847498.
16. Chiew CJ, Liu N, Tagami T, Wong TH, Koh ZX, Ong MEH. Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department. Medicine (Baltimore). 2019;98(6):e14197. Epub 2019/02/09. doi: 10.1097/MD.0000000000014197. PubMed PMID: 30732136; PubMed Central PMCID: PMCPMC6380871.
17. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med. 2016;23(3):269-78. Epub 2015/12/19. doi: 10.1111/acem.12876. PubMed PMID: 26679719; PubMed Central PMCID: PMCPMC5884101.
18. Perng JW, Kao IH, Kung CT, Hung SC, Lai YH, Su CM. Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning. J Clin Med. 2019;8(11). Epub 2019/11/11. doi: 10.3390/jcm8111906. PubMed PMID: 31703390; PubMed Central PMCID: PMCPMC6912277.
19. Fagerstrom J, Bang M, Wilhelms D, Chew MS. LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock. Sci Rep. 2019;9(1):15132. Epub 2019/10/24. doi: 10.1038/s41598-019-51219-4. PubMed PMID: 31641162; PubMed Central PMCID: PMCPMC6805937.
20. Mao Q, Jay M, Hoffman JL, Calvert J, Barton C, Shimabukuro D, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8(1):e017833. Epub 2018/01/29. doi: 10.1136/bmjopen-2017-017833. PubMed PMID: 29374661; PubMed Central PMCID: PMCPMC5829820.
21. Klug M, Barash Y, Bechler S, Resheff YS, Tron T, Ironi A, et al. A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score. J Gen Intern Med. 2020;35(1):220-7. Epub 2019/11/05. doi: 10.1007/s11606-019-05512-7. PubMed PMID: 31677104.
22. Sahni N, Simon G, Arora R. Development and Validation of Machine Learning Models for Prediction of 1-Year Mortality Utilizing Electronic Medical Record Data Available at the End of Hospitalization in Multicondition Patients: a Proof-of-Concept Study. J Gen Intern Med. 2018;33(6):921-8. Epub 2018/02/01. doi: 10.1007/s11606-018-4316-y. PubMed PMID: 29383551; PubMed Central PMCID: PMCPMC5975145.
23. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One. 2017;12(4):e0174708. Epub 2017/04/07. doi: 10.1371/journal.pone.0174708. PubMed PMID: 28384212; PubMed Central PMCID: PMCPMC5383046.
24. Ford DW, Goodwin AJ, Simpson AN, Johnson E, Nadig N, Simpson KN. A Severe Sepsis Mortality Prediction Model and Score for Use With Administrative Data. Crit Care Med. 2016;44(2):319-27. Epub 2015/10/27. doi: 10.1097/CCM.0000000000001392. PubMed PMID: 26496452; PubMed Central PMCID: PMCPMC4724863.
25. Shukeri W, Ralib AM, Abdulah NZ, Mat-Nor MB. Sepsis mortality score for the prediction of mortality in septic patients. J Crit Care. 2018;43:163-8. Epub 2017/09/14. doi: 10.1016/j.jcrc.2017.09.009. PubMed PMID: 28903084.
26. Bogle B, Balduino, Wolk D, Farag H, Kethireddy, Chatterjee, et al. Predicting Mortality of Sepsis Patients in a Multi-Site Healthcare System using Supervised Machine Learning2019.
27. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv e-prints [Internet]. 2016 March 01, 2016. Available from: https://ui.adsabs.harvard.edu/abs/2016arXiv160302754C.
28. Nanayakkara S, Fogarty S, Tremeer M, Ross K, Richards B, Bergmeir C, et al. Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLoS Med. 2018;15(11):e1002709. Epub 2018/12/01. doi: 10.1371/journal.pmed.1002709. PubMed PMID: 30500816; PubMed Central PMCID: PMCPMC6267953 following competing interests: KR is director of IntelliHQ Pty Ltd, non-profit AI innovation centre for healthcare, connected with Gold Coast University Hospital. KR is owner and Chairman of K. J. Ross & Associates Pty. Ltd. (KJR), professional services firm specialising in IT risk management and assurance. 20% of KJR's work is in healthcare. There is no direct financial stake in the results of the current study.

29. Levy MM, Fink MP, Marshall JC, Abraham E, Angus D, Cook D, et al. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. Crit Care Med. 2003;31(4):1250-6. Epub 2003/04/12. doi: 10.1097/01.CCM.0000050454.01978.3B. PubMed PMID: 12682500.
30. World Medical A. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191-4. Epub 2013/10/22. doi: 10.1001/jama.2013.281053. PubMed PMID: 24141714.
31. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749-60. Epub 2019/04/20. doi: 10.1038/s41551-018-0304-0. PubMed PMID: 31001455; PubMed Central PMCID: PMCPMC6467492.
32. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. Explainable AI for Trees: From Local Explanations to Global Understanding. arXiv e-prints [Internet]. 2019 May 01, 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190504610L.
33. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020. doi: 10.1038/s42256-019-0138-9.
34. Lipovetsky S, Conklin M. Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry. 2001;17(4):319-30. doi: 10.1002/asmb.446.
35. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems. 2013;41:647-65.
36. Chang SH, Hsieh CH, Weng YM, Hsieh MS, Goh ZNL, Chen HY, et al. Performance Assessment of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score, and Rapid Acute Physiology Score in Predicting Survival Outcomes of Adult Renal Abscess Patients in the Emergency Department. Biomed Res Int. 2018;2018:6983568. Epub 2018/10/18. doi: 10.1155/2018/6983568. PubMed PMID: 30327779; PubMed Central PMCID: PMCPMC6169207.
37. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. Proceedings of the 22nd international conference on Machine learning; Bonn, Germany. 1102430: ACM; 2005. p. 625-32.
38. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review. 1950;78(1):1-3. doi: 10.1175/1520-0493(1950)078<0001:Vofeit>2.0.Co;2.
39. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45. Epub 1988/09/01. PubMed PMID: 3203132.
40. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. Epub 2012/10/25. PubMed PMID: 23092060; PubMed Central PMCID: PMCPMC3900052.
41. Shimabukuro DW, Barton CW, Feldman MD, Mataraso SJ, Das R. Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial. BMJ Open Respir Res. 2017;4(1):e000234. Epub 2018/02/13. doi: 10.1136/bmjresp-2017-000234. PubMed PMID: 29435343; PubMed Central PMCID: PMCPMC5687546.

Supplemental material

Supporting information

S1 supporting information. Extended description of clinical criteria and risk scores.

In the current manuscript we describe several clinical criteria and risk scores. Below is a detailed description of each score or criteria and their application in our manuscript:

- qSOFA: the simple quick sequential organ failure assessment (qSOFA) is a bedside tool to identify patients with suspected infection who are at greater risk for a poor outcome. It uses three criteria, assigning one point for low blood pressure (systolic blood pressure ≤ 100 mmHg), high respiratory rate (≥22 breaths per min), or altered mentation (Glasgow coma scale < 15). In the current study we included patients with ≥ 2 points.
- SIRS: the systemic inflammatory response syndrome (SIRS) criteria has the same objective as the qSOFA criteria. SIRS assigns one point for tachycardia (heart rate > 90 beats/min), tachypnea (respiratory rate > 20 breaths/min), fever or hypothermia (temperature >38 or <36 °C), and leukocytosis, leukopenia, or bandemia (white blood cells > 12 * 109/L , < 4 * 109/L or bandemia ≥ 10%). In the current study we included patients with ≥2 points.
- abbMEDS: the abbreviated Mortality Emergency Department Sepsis (abbMEDS) score assesses sepsis severity and predicts mortality. This score assigns six points for terminal disease, three for respiratory difficulty (respiratory rate >30 breaths/min), three for septic shock, three for low thrombocytes (< 150 * 109/L), three for a higher age (> 65 years), two for a lower respiratory tract infection, two for being a nursing home resident and two for an altered mental state (Glasgow coma scale < 15). abbMEDS is used in three categories: low (0 – 4 points), intermediate (5 – 12 points) and high risk (13 – 24), we used a cut-off value of 7 to dichotomize the outcome.
- mREMS: the modified rapid emergency medicine score (mREMS) is a triage score that is used to predict 28-day in-hospital mortality. This score assigns points for age (0: ≤44, 1: 45-64, 3: 65-74, 4: >74), systolic blood pressure (0: 110-159, 1: 160-199 and 90-109, 2: ≥200 and 80-89, 4: ≤79), heart rate (0: 70-109, 2: 110-139 and 55-69, 3: 140-179 and 40-54, 4: >179 and ≤39), respiratory rate (0: 12-24, 1: 25-34 and 10-11, 2: 6-9, 3: 35-49, 4: >49 and ≤5), oxygen saturation (0: ≥89, 1: 86-89, 3: 75-85, 4: <75) and the Glasgow coma scale (0: 14 or 15, 2: 8-13, 5: 5-7, 6: 3 or 4).
- SOFA: the SOFA-score was initially developed as a tool to learn from the evolution of organ failure in sepsis, but later was extensively validated to predict morbidity and mortality in several populations (Ceriani et al., 2003, Chest; Minne et al., 2008, Crit Care). It scores 1-4 points for each of the six organ systems, and we calculated it according to the formula described in the original paper (Vincent et al., 1996, Inte Care Med).

S2 supporting information. Background information on machine learning models reviewed in the current study.

We conducted a comparison of available algorithms on the 31-day mortality prediction task. We considered the following algorithms:

- **Logistic regression:** logistic regression is a statistical technique that in its basic form uses a logistic function to model a binary dependent variable. A simple logistic regression model was used with L2 regularization, a tolerance of 1e-4 and a maximum amount of iterations of 1,000. We used the limited Broyden–Fletcher–Goldfarb–Shanno (lbfgs) algorithm as optimizer. Logistic regression was implemented using Python (version 3.7.1) and the sklearn (version 0.22.1) package.
- **Multi-layer perceptron neural network:** neural networks are statistical models vaguely inspired by the biological neural networks that constitute animal brains. A simple feed-forward, multi-layer perceptron neural network was implemented consisting of three hidden layers with respectively 128, 64 and 32 neurons and ReLU activation functions. We trained the network using a constant learning rate of 0.001 with a batch size of 1 and the Adam optimization scheme. The neural network was implemented using Python programming language (version 3.7.1) using packages Keras (version 2.2.2) and scikit-learn (version 0.22.1).
- **Random Forest:** random forests is an ensemble learning method for classification that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).A random forest classifier with a decision tree as base learner, consisting of 200 trees with gini criterion and a maximum depth of 50 was used. We used bootstrapped samples for building trees.
- **Gradient-boosting systems:** gradient boosting is a machine learning technique for classification problems which produce a prediction model in the form of an ensemble of weak prediction models, typically decision trees. In contrast to random forests, it builds the model in a stage-wise fashion like other boosting methods do. We used the XGBoost implementation of gradient-boosting systems. Each implementation has specific unique implementation details, but all use decision trees as the base weak learner and gradient boosting to iteratively fit a sequence of such trees. We used a learning rate of 0.075, a maximum number of trees of 300 and a maximum depth of each base learner to be 13. We implemented this using the Python programming language (version 3.7.1) using the package XGBoost (version 0.90).

Supporting Tables

S1 Table. Overview of variables present in the datasets described in the manuscript.
The laboratory dataset consisted exclusively of laboratory variables with age, sex , and time of request. The laboratory and clinical dataset contained all variables from the laboratory dataset and additionally clinical and vital characteristics.

Laboratory dataset[1]		**Laboratory and clinical dataset**
Age	HDL	Laboratory dataset variables
Sex	Hematocrit	Weight
Request time	Hemoglobin	Length
Sodium bicarbonate	INR	Policy restrictions
ALAT	Iron	Respiratory rate
Albumin	Lactate	Saturation
Alkaline phosphatase	Lactate dehydrogenase	Temperature
Alpha-1-fetoprotein	Leukocytes	Glasgow coma score
Ammonia	Lipase	FIO2
Amylase	Lymphocytes	Heart rate
Anti Xa	Magnesium	Systolic BP
Anti-thrombin	MCV	Diastolic BP
APTT	MDRD	Inotropics/vasopressors
ASAT	Metamyelocyte	ECG rhythm
Atypical lymphocytes	Monocyte	
Base excess	Myelocyte	
Basophiles	Neutrophils	
Bilirubin	NT-proBNP	
Blasts	Osmolality	
Blood transfusion	pCO2	
Calcium ion	pH	
Calcium total	Phosphate	
Chloride	Platelet count	
CK	pO2	
CKD-EPI	Poikilocytosis	
CK-MB	Potassium	
Cortisol	Promyelocytes	
Creatinin	PT	
CRP	PTH	
D-Dimers	Reticulocyte	
Direct Antiglobulin Test	Rod-like granulocytes	
Dysmorphic erytrocytes	Sedimentation rate	
Eosinophil	Segment core granulocytes	
Erythroblasts	Sodium	
Erythrocytes	Specific gravity	
Estradiol	Standard sodium bicarbonate	
Ferritin	Total CO2	
Fibrinogen	Total protein	
Folic acid	Toxic grain	
Fragmentocytes	Transferrin	
PSA	Transferrin saturation	
Free T4	Triglycerides	
Gamma GT	Troponin T	
Gentamycin	TSH	
Glucose	Urea	
Haptoglobin	Uric acid	
HbCO	Urobilinogen	
HbO2	Vitamin 25 (OH) D3	
	Vitamin B12	

[1] *For each of the laboratory variables we generated an additional binary 'absence' or 'presence' variable representing whether or not this laboratory parameter was requested.*

S2 Table. Comparison of baseline statistical and machine learning models for predicting 31-day mortality risk.

We performed a baseline comparison of statistical and machine learning models (S1 supporting information) for the 31-day mortality prediction task using the laboratory dataset. We used five-fold cross validation to assess model performance. Performance was assessed by area under the receiver operating characteristic curve (AUC) and accuracy.

Evaluation metric	Logistic regression	Multi-layer perceptron	Random Forest	XGBoost
AUC	0.633 (0.606 – 0.660)	0.658 (0.632 – 0.685)	0.723 (0.689 – 0.756)	**0.813 (0.791 – 0.835)**
Accuracy	0.826 (0.820 – 0.833)	0.868 (0.858 – 0.877)	0.842 (0.831 – 0.853)	**0.873 (0.864 – 0.883)**

S3 Table. Hyperparameters of XGBoost models.
Hyperparameters were based on theoretical reasoning rather than hyperparameter tuning. This was done to prevent overfitting on hyperparameters due to small sample size. "Base_score", "Missing", "Reg_alpha", "Reg_lambda" and "Subsample" parameters were standard values provided by the XGBoost interface. "Max_depth", "max_delta_step" and "estimators" were values we internally use for these kind of machine learning models. During the study, hyperparameters were never adjusted to gain performance in our validation dataset.

Hyperparameter	Value	Explanation
Max_depth	13	Determines how deeply each tree is allowed to grow during any boosting round.
Max_delta_step	3	Maximum delta step we allow each tree's weight estimation to be.
Learning rate	0.075	Degree to which weights are adjusted each learning iteration.
Base_score	0.5	Initial prediction score of all instances (global bias).
Missing	N/A	Value which is represented as missing. Put onto N/A as no imputation was performed in data processing.
Reg_alpha	0	L1 regularization term on weights
Reg_lambda	1	L2 regularization term on weights
Subsample	1	Percentage of samples used per tree; low value can lead to underfitting.
Estimators	300	Number of trees you want to build.

S4 Table. Extended analysis of correlation between important model features and clinical risk scores. To study the correlation between the most important features contributing to model predictions and the clinical criteria (qSOFA and SIRS) and risk scores (abbMEDS and mREMS), we compared their existence in both. The top-20 most important features (Fig. 3 in manuscript) are compared to all criteria in the clinical scores (S1 supporting information). Most of the features present in the clinical criteria and scores are also among the most important features in the lab + clinical machine learning model.

Lab model top-20 features	Clinical criteria and scores				Lab/ clinical model top-20 features	Clinical criteria and scores			
	SIRS 0/4	qSOFA 0/3	MEDS 2/6	REMS 1/6		SIRS 3/4	qSOFA 3/3	MEDS 4/6	REMS 5/6
1. Blood group (ordered)					1. Heart rate	X			X
2. Urea					2.Blood group (ordered)				
3. Albumin					3. Urea				
4. CKD-EPI					4. Albumin				
5. Blood group					5. Magnesium				
6. Platelet count			X		6. Platelet count			X	
7. Age			X	X	7. GCS		X	X	X
8. Total protein					8. Oxygen saturation	X	X	X	X
9. CRP					9. Age			X	X
10. Sodium					10. Temperature	X			
11. Glucose (arterial)					11. Glucose				
12. LD					12. Systolic BP		X		X
13. Lactacte					13. Creatinine				
14. Creatinine					14. Amount of lab				
15. Lipase					15. CRP				
16. Bilirubin					16. ALAT				
17. Gamma-GT					17. A. Fib (history)				
18. Alk. phosphatase					18. CO2 (arterial)				
19. Magnesium					19. Bilirubin				
20. Hemoglobin					20. Calcium				

S5 Table. Extended comparison of machine learning model with internal medicine physicians and clinical risk scores.

In addition to sensitivity and specificity, we evaluated the performance of each group by positive predictive value (PPV), negative predictive value (NPV), accuracy and area-under-the receiver operating characteristics curve (AUC). The machine learning model shows superior performance in each of these metrics, which is consistent with the findings presented in the manuscript. Confidence intervals were calculated using binomial testing and AUC's were compared using DeLong's test.

Evaluation metric	XGBoost model	abbMEDS	mREMS	SOFA	Internal medicine physicians
Sensitivity, %	92.3 (87.1 – 95.3)	53.8 (44.1 – 63.6)	61.5 (52.0 – 71.1)	76.9 (68.7 – 85.2)	72.1 (61.3 – 82.2)
Specificity, %	78.2 (70.1 – 86.3)	72.4 (63.7 – 81.2)	64.4 (55.0 – 73.8)	73.6 (64.9 – 82.2)	74.2 (63.9 – 82.1)
PPV, %	38.7 (29.2 – 48.3)	22.6 (14.4 – 30.8)	20.5 (12.6 – 28.4)	30.3 (21.3 – 39.3)	29.5 (20.5 – 38.4)
NPV, %	98.6 (96.2 – 100.0)	91.3 (85.8 – 96.8)	91.8 (86.4 – 97.2)	95.5 (91.5 – 99.6)	94.8 (90.5 – 99.2)
Accuracy	0.800 (0.722 – 0.878)	0.700 (0.610 – 0.790)	0.640 (0.546 – 0.734)	0.740 (0.654 – 0.826)	0.738 (0.651 – 0.824)
AUC	0.852 (0.783 – 0.922)	0.631 (0.537 – 0.726)	0.630 (0.535 – 0.724)	0.752 (0.667 – 0.836)	0.735 (0.648 – 0.821)
P-value	N/A	0.021	0.016	0.042	0.189; 0.072; 0.068; 0.032[a]

[a] *Individual P-values were calculated for each of the internal medicine physicians.*

S6 Table. Inter-rater agreement of internal medicine physicians.

Cohen's kappa was used to measure the inter-rater agreement between the internal medicine physicians. The level of agreement was interpreted as nil if κ was 0 to 0.20; minimal, 0.21 to 0.39; weak, 0.40 to 0.59; moderate, 0.60 to 0.79; strong, 0.80 to 0.90; and almost perfect, 0.90 to 1.0.

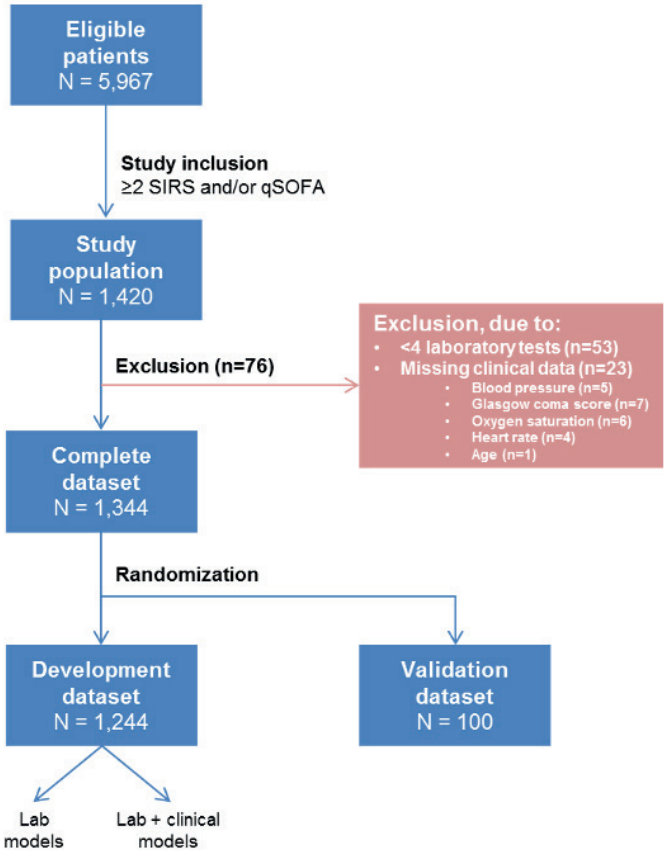
Internist	1 (consultant)	2 (consultant)	3 (resident)	4 (resident)
1 (consultant)	-	0.52	0.54	0.46
2 (consultant)	0.52	-	0.61	0.67
3 (resident)	0.54	0.61	-	0.63
4 (resident)	0.46	0.67	0.63	-

S7 Table. Machine learning comparison to alternating physician groups.
In each comparison between the machine learning model and the physicians group, a single physician was removed from the physician group. In every comparison the machine learning model outperforms the physicians. This analysis shows that the higher performance of the machine learning model was not due to systemic underperformance of a single physician.

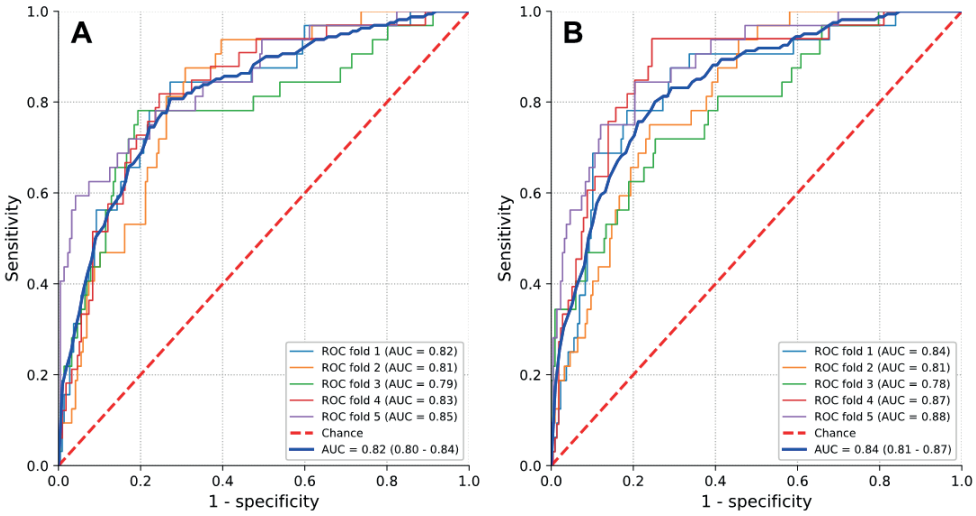
Group	Sensitivity		Specificity	
	Mean (95% CI)	P-value compared to model	Mean (95% CI)	P-value compared to model
All internists	0.72 [0.62-0.81]	<0.001	0.74 [0.64-0.82]	0.509
Internist 2, 3, 4	0.71 [0.61-0.80]	<0.001	0.75 [0.65-0.83]	0.524
Internist 1, 3, 4	0.69 [0.60-0.78]	<0.001	0.76 [0.67-0.84]	0.653
Internist 1, 2, 4	0.74 [0.66-0.83]	0.001	0.74 [0.65-0.82]	0.447
Internist 1, 2, 3	0.77 [0.69 – 0.85]	0.003	0.72 [0.63-0.81]	0.316

Supporting Figures

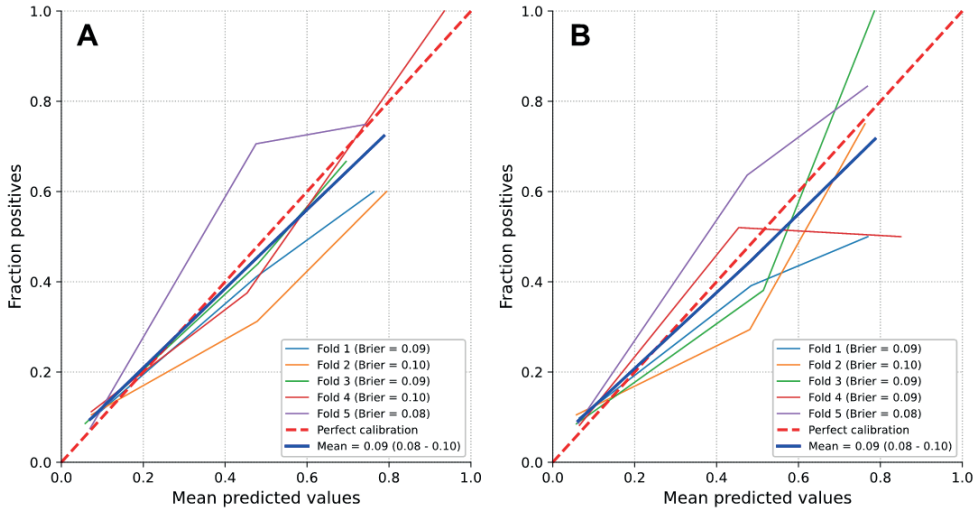
S1 Fig. Flow diagram of study inclusion.
During the study period 5,967 patients that presented to our emergency department were referred to an internal medicine physician. Of these patients, 1420 patients fulfilled two or more SIRS and/or qSOFA criteria. After exclusion of 76 patients, a number of 1,344 patients were separated into development and validation datasets.



S2 Fig. Five-fold cross validation of diagnostic performance of XGBoost models.
During each fold of cross-validation, we assessed predictive performance by area under the receiver operating characteristic curves (AUC). Performance was determined for models trained with laboratory data (A) and models trained with laboratory + clinical data (B) to predict 31-day mortality.

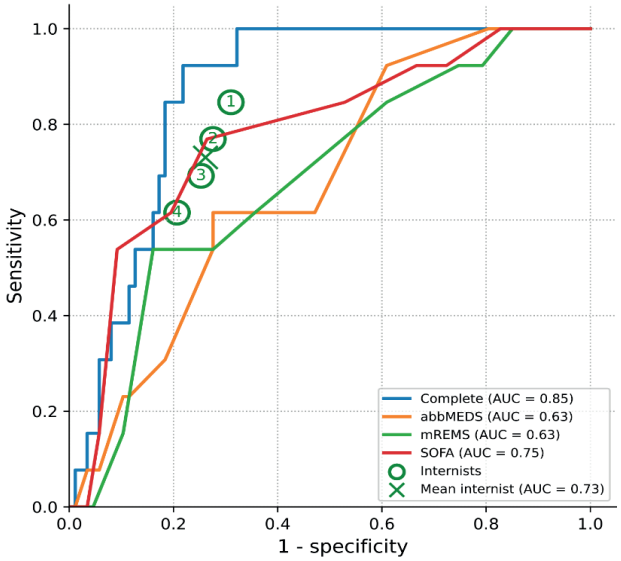


S3 Fig. Five-fold cross validation of calibration of XGBoost models.
During each fold of cross-validation, we assessed calibration by calibration curves and their respective brier scores. Calibration was determined for models trained with laboratory data (A) and models trained with laboratory + clinical data (B).

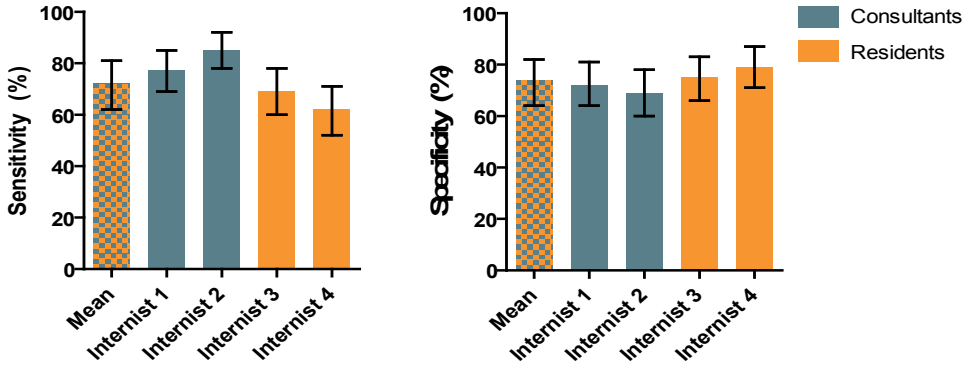


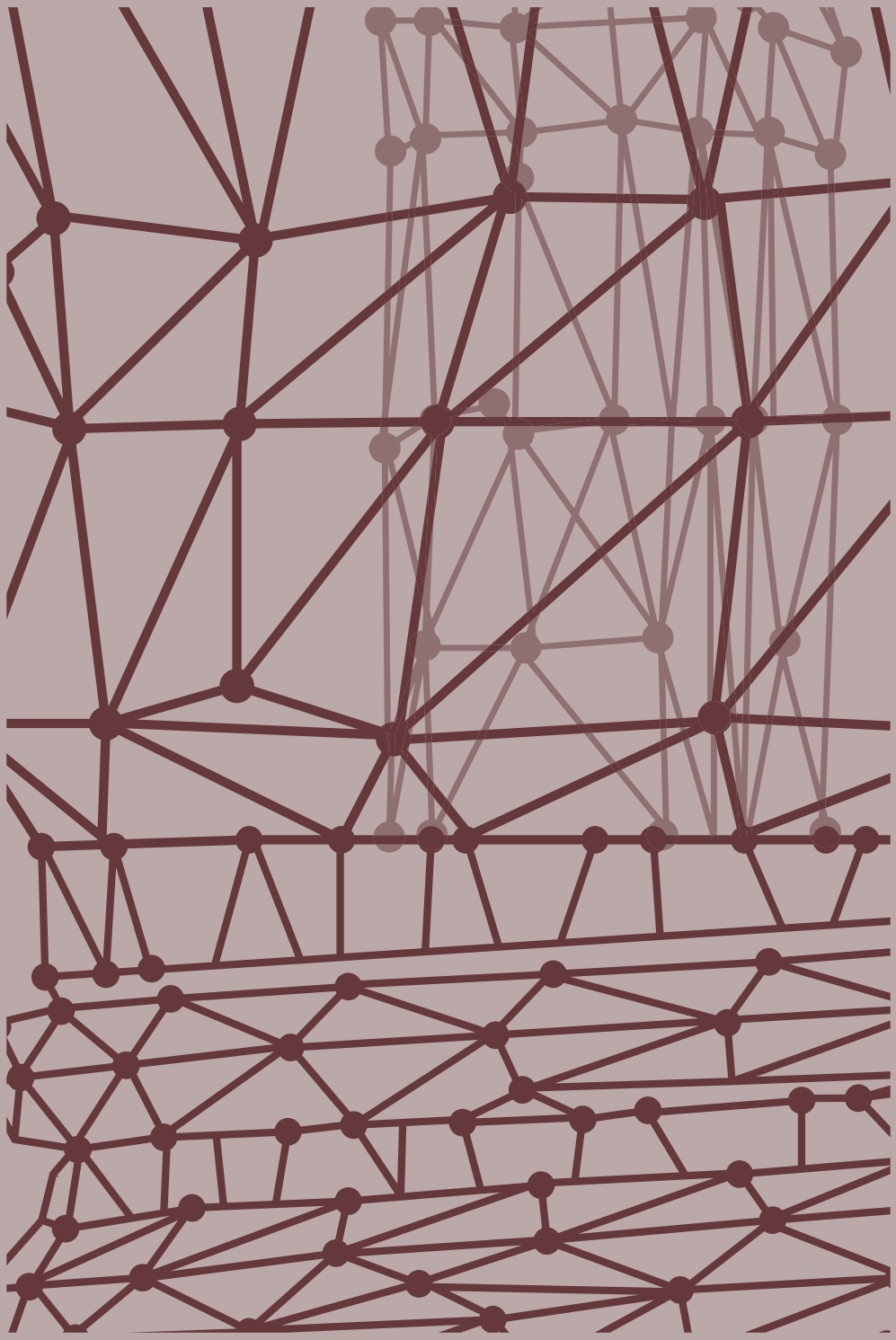
S4 Fig. Receiver operating characteristic analysis of machine learning model, risk scores and internal medicine physicians.

Receiver operating characteristics analysis of the lab + clinical machine learning model (AUC: 0.852 [0.783-0.922]), abbMEDS (0.631 [0.537-0.726]), mREMS (0.630 [0.535-0.724]), SOFA (AUC: 0.752 [0.667 – 0.836]) and internal medicine physicians (mean 0.735 [0.648-0.821]). Internal medicine physicians were depicted as bullets in the ROC analysis.



S5 Fig. Individual performance of internal medicine physicians.
Predictive performance of all internal medicine specialists (n=4; 2 experienced consultants in acute internal medicine and 2 experienced residents acute internal medicine) was assessed by sensitivity (left) and specificity (right). Consultants are depicted in grey and residents in orange.





CHAPTER 8

EXPLAINABLE MACHINE LEARNING MODELS FOR RAPID RISK STRATIFICATION IN THE EMERGENCY DEPARTMENT: A MULTI-CENTER STUDY

William P.T.M. van Doorn, Floris Helmich, Paul M.E.L. van Dam,
Leo H.J. Jacobs, Patricia M. Stassen, Otto Bekers, Steven J.R. Meex

SUBMITTED FOR PUBLICATION

Abstract

Introduction**:** Risk stratification of patients presenting to the emergency department (ED) is important for appropriate triage. Using machine learning crude laboratory data requested at these EDs can be modelled to improve the risk stratification of individual patients. In this study, we aimed to apply machine learning to develop an accurate and explainable clinical decision support tool model that predicts the likelihood of 31-day mortality in ED patients (the RISK^{INDEX}). This tool was developed and evaluated in 4 Dutch hospitals.

Methods: Machine learning models included patient characteristics and available laboratory data collected within the first two hours after ED presentation, and were trained using five years of data from consecutive ED patients from the Maastricht University Medical Centre+ (Maastricht), Meander Medical Center (Amersfoort), and Zuyderland (locations Sittard and Heerlen). A sixth year of data from ED patients was used to evaluate the models using area under the receiver-operating-characteristic curve (AUROC), brier scores and calibration curves. The SHapley Additive exPlanations (SHAP) algorithm was used to obtain explainable machine learning models.

Results: The present study included 266,327 patients with more than 7 million laboratory results available. Models showed high diagnostic performance with AUROCs of 0.94 [0.94-0.95], 0.98 [0.97-0.98], 0.88 [0.87-0.89] and 0.90 [0.89-0.91] for Maastricht, Amersfoort, Sittard and Heerlen, respectively. The SHAP algorithm was utilized to visualize patient characteristics and laboratory data patterns that underlie individual RISK^{INDEX} predictions.

Discussion: Novel clinical decision support tools based on an explainable machine learning model have excellent diagnostic performance in predicting 31-day mortality in ED patients across four hospitals. Follow-up studies will assess whether implementation of these algorithm can improve clinically relevant endpoints.

Introduction

An increasing number of patients are referred to emergency departments (ED) worldwide [1-3]. Prolonged waiting times and associated crowding in the ED increase mortality, [4] and rapid risk stratification is therefore a core task in emergency medicine. An effective means to identify patients at high- and low-risk shortly after admission could help decision-making regarding patient prioritization, treatment, level of observation, and post-discharge follow-up. Consequently, numerous clinical risk scores and triage systems for stratification of patients in the ED have been developed, such as the modified early warning score (MEWS), rapid emergency medicine score (REMS) and emergency severity index (ESI) [5-9]. Unfortunately, these systems often generalize poorly and lack precision, impeding their clinical use [10, 11].

Emergency departments generate vast amounts of clinical, physical and laboratory data. This data is generally heterogeneous and comprises both structured and unstructured information. Machine learning allows processing and modelling of this data to a human interpretable level in relation to clinically relevant endpoints. Accordingly, machine-learning based mortality prediction models were developed using data extracted from patients in the ED [12-19]. Although these models were superior to traditional risk scores and physicians [12, 15, 17, 20], most are perceived as so-called “black boxes”, possibly limiting their acceptance among clinicians and raising legal or ethical concerns. Recently, models that are transparent in their patient-specific risk predictions have emerged [21-24]. This development may not only advance our understanding machine learning based algorithms, but also contribute to more widespread acceptance of clinical decision-support tools based on machine learning technology among clinicians.

In this study, our aim was to develop an accurate clinical decision support tool in four hospitals in The Netherlands using machine learning technology. These models were designed to combine patient characteristics and early available laboratory results at the ED to generate an individuals’ likelihood of 31-day mortality: the RISKINDEX.

Methods

Study design and setting

A multi-center, retrospective cohort study was performed among all patients who presented to the ED at the Maastricht University Medical Center (Maastricht, The Netherlands), Meander Medical Center (Amersfoort, The Netherlands) and Zuyderland medical Center locations Sittard (Sittard, The Netherlands) and Heerlen (Heerlen, The Netherlands) between January 1, 2013 and December 31, 2018. For convenience, each of the centers will be referred to by their respective location; Maastricht, Amersfoort, Sittard and Heerlen. This study was approved by the medical ethical committees of each of the individual centers (Maastricht: #2018-0838, Amersfoort: TWO19-46, Sittard: #2018-0838, Heerlen: #2018-0838). The study follows the STROBE guidelines [25] and was conducted according to the principles of the Declaration of Helsinki [26].

Patient population

All patients presenting to the ED aged ≥18 years with at least 3 laboratory tests ordered by the attending physician were included. Patients whose previous presentation to the ED was less than 48 hours ago were excluded.

Dataset construction

Data anonymization, collection, processing, model selection, development and evaluation were performed for each of the four hospitals separately. All available laboratory data of the patients ordered within two hours after the first laboratory request from the ED were collected. All laboratory data acquired after two hours were not used for model development. In addition, rare laboratory tests (<0.01%) were excluded. By restricting the parameters in the model to laboratory tests ordered within the first two hours after the first test, we aimed to develop a decision support tool that would allow rapid triage after presentation. The primary outcome measure for the study was mortality within 31 days after initial ED presentation and was acquired through electronic health records.

For each hospital six years of data from consecutive patients were available. Data from the first five years was used for model development, and data from the sixth year was used to validate performance of each model. The model development data was randomly split into training (70%), tuning (20%) and calibration datasets (10%) such that data from a given presentation was present in one split only. Consequently, the training split was used to train the proposed models, the tuning set was used to iteratively improve the models by selecting the best model architectures and hyperparameters, and the calibration split was used to perform post-hoc calibration on the model predictions. Finally, the validation dataset (consisting of year six data), was used to evaluate the performance of machine learning models.

Model selection, training and calibration

The clinical decision support tool combines patient characteristics and laboratory results through machine learning to predict the likelihood of 31-day mortality. The output of the clinical decision support tool -termed the RISKINDEX- is a calibrated value between 0 and 100. Various statistical and machine learning algorithms can be applied to develop such a clinical decision support tool, including regression techniques [27, 28], neural network architectures [29, 30], gradient boosting systems [31-34] and decision trees [35] (Supplemental section A).

The light gradient boosting system (LightGBM) architecture was selected amongst several alternatives on the basis of the tuning set performance (Supplemental information section A and Table 1). LightGBM is an implementation of distributed, efficient gradient-boosting systems with native support for missing values [33]. Consequently, a broad spectrum of hyperparameter combinations for this architecture was evaluated (Supplemental Table 2). Hyperparameter optimization is the process of selecting a set of optimal hyperparameters, which are features controlling the training process of a machine learning model; such as the rate of learning and the maximum level of complexity. In the current study, bayesian hyperparameter optimization using tree-parzen estimators (TPE) was utilized [36]. This approach is based on building a probability model of the objective function and using this to select the most promising hyperparameters to evaluate in the true objective function. Optimization was run for 1,000 iterations with logarithmic loss as our objective function (Supplemental Table 2 lists the definitions of the search space per hyperparameter). Hyperparameter optimization resulted in LightGBM architectures consisting of 220 – 740 boosted trees with a maximum depth of 11 – 37 and maximum leaves of 320 – 690 for each base learner (see Supplemental Table 3). Exponential learning-rate decay was used during training with an initial learning rate of 0.075 – 0.145 decaying every 2,000 training steps by a factor of 0.7-0.8. The loss function during training was logarithmic loss.

LightGBM models with optimal hyperparameters were recalibrated on the calibration set in order to further improve the quality of the generated RISKINDEX. Recalibration is recommended as most LightGBM models are prone to miscalibration, meaning that their output RISKINDEX do not represent the actual 31-day mortality likelihood. Hence, recalibration ensures that consistent probabilistic interpretations of the RISKINDEX predictions can be made [37]. For calibration, various techniques were considered including Platt scaling [38], isotonic regression [39] and Platt-Binner scaling [40]. Model calibration was assessed by the brier score [41] and visual inspection of reliability plots [42]. Reliability plots are the usual approach for evaluating calibration of binary outcomes in which we compare decile-binned means of predictions versus means of the observed outcomes

in the patients. Platt-Binner scaling was selected as this was shown to result in the best calibrated models (see Supplemental Figure 1). The resulting calibrated predictions were defined as the RISKINDEX. Data preprocessing, model development, selection, training and calibration was performed using Python programming language (version 3.7.1) using packages Numpy (version 1.17), Pandas (version 0.24), Keras (version 2.2.2), scikit-learn (version 0.22.0) and tensorflow (version 2.0.1, beta).

Model evaluation

Overall model performance was evaluated in the validation set of each hospital separately, all of them containing 1 full year of data from ED patients, not previously used for model development. Evaluation was done by 1) area under the receiver-operating-characteristic curve (AUROC) to quantify the ability of models to discriminate between survivors and non-survivors, and 2) visual inspection of calibration curve and Brier scores to estimate how accurately the RISKINDEX estimates the likelihood of 31-day mortality. Next, an embedded reference table was created based on the validation dataset to report estimates of sensitivity, negative predictive value (NPV), specificity and positive predictive value (PPV) for each RISKINDEX between 0-100. This table was subsequently used to compare diagnostic metrics from the model (sensitivity, NPV, specificity, and PPV) across the four hospitals at various selected statistical thresholds.

Model explanation

To facilitate the interpretation of the RISKINDEX generated by our machine learning models, the Shapley additive explanations (SHAP) algorithm was applied [21, 43]. SHAP highlights the patient characteristics and laboratory results (further referred to as "variables") that underlie patient-specific predictions by the model, hence mitigating the issue of black-box predictions. SHAP is a model-agnostic representation of feature importance where the impact of each variable on a particular prediction is represented using Shapley values inspired by cooperative game theory and their extensions [44-46]. A Shapley value states -given the current set of variables- how much a variable in the context of its interaction with other variables contributes to the difference between the actual prediction and the mean prediction. That is, the mean prediction plus the sum of the Shapley values for all variables equals the actual prediction. It is critically important to understand that this is fundamentally different to direct variable effects known from e.g. (generalized) linear models. The SHAP value for a variable should not be seen as its direct -and isolated effect- but as its aggregated effect when interacting with other variables in the model. In our specific case, positive Shapley values contribute towards a positive prediction (death), whilst low or negative Shapely values contribute towards a negative prediction (survival).

Statistical analysis

Descriptive analysis of baseline characteristics was performed using IBM SPSS Statistics for Windows (version 24.0). Continuous variables were reported as means with standard deviation (SD) or medians with interquartile ranges (IQRs) depending on the distribution of the data. Categorical variables were reported as proportions. Thousand bootstrap iterations were used to calculate 95% confidence intervals, unless otherwise mentioned. Model evaluation and statistical analysis was performed using Python (version 3.7.1) using packages Numpy (version 1.17), Pandas (version 0.24) and Matplotlib (version 3.1.2).

Results

Patient and laboratory characteristics

In the current study more than 50,000 presentations were included for each hospital resulting in a total of 266,327 unique presentations to the ED. The total population consisted of slightly more female (mean; 50.8%) patients with a mean age of 61.5 (+- 22.4) years. Within the first two hours of presentations, on average 29 (± 10.6) laboratory parameters were requested. Although requested parameters were subject to intercenter heterogeneity, complete blood count, electrolytes and lactate dehydrogenase (LD) represented the most prevalent in all hospitals. Mortality rates at 31 days were 6.4%, 4.0%, 5.9% and 5.0% for Maastricht, Amersfoort, Sittard and Heerlen, respectively. Baseline characteristics are described in Table 1.

Table 1. Baseline patient and laboratory characteristics of the four study populations.

	MUMC, Maastricht N = 66,770	**Meander, Amersfoort N = 81,152**	**Zuyderland, Sittard N = 66,423**	**Zuyderland, Heerlen N = 51,982**
Demographics				
Age, years	59 (± 22.2)	59 (± 23.7)	65 (± 21.3)	64 (± 21.5)
Sex, female (%)	32,829 (49.2%)	41,907 (51.6%)	34,256 (51.6%)	26,185 (50.3%)
Laboratory				
Mean number of tests per patient	22 (± 10.0)	25 (± 8.0)	31 (± 10.4)	31 (± 10.8)
Ten most frequent laboratory orders, n (%)	Creatinine, 59,937 (89.9%) CBC[a], 59,632 (89.4%) Sodium, 56,529 (84.8%), Potassium, 56,487 (84.7%) CRP[b], 55,178 (82.7%) Urea, 53,972 (80.9%) Platelets, 47,059 (70.6%) Glucose, 43,733 (65.6%) ALAT[c], 36.429 (54.6%) ASAT[d], 32.130 (48.1%)	CBC[a], 77,625 (95.7%) CRP[b], 76,639 (94.4%) Sodium, 75,638 (93.2%) Creatinin, 75,393 (92.9%) Potassium, 75,025 (92.4%) Glucose, 75,018 (92.4%) Urea, 72.115 (88.9%) CK, 64.356 (79.3%) Platelets, 54.380 (67.0%) ALAT[c], 54.279 (66.9%)	CBC[a], 62,518 (94.1%) Platelets, 62,517 (94.1%) CRP[b], 60,910 (91.7%) Creatinin, 60,046 (90.4%) Sodium, 58,558 (88.1%) Potassium, 58,404 (87.9%) Glucose, 57,987 (87.3%) Urea, 57,559 (86.6%) ALAT[c], 53,968 (81.2%) ASAT[d], 53.914 (81.2%)	CBC[a], 49,056 (94.4%) Platelets, 49,040 (94.3%) CRP[c], 47,573 (91.5%) Creatinin, 47,043 (90.5%) Sodium, 46,615 (89.7%) Potassium, 46,364 (89.2%) Glucose, 45,536 (87.6%) Urea, 45,074 (86.7%) ALAT[c], 42,486 (81.7%) ASAT[d], 42.415 (81.6%)
Outcome				
31-day mortality	4,242 (6.4%)	3,277 (4.0%)	3,917 (5.9%)	2,603 (5.0%)

Model performance

Using the available patient characteristics and laboratory data from the first two hours of a presentation, we developed machine learning models that predict the 31-day mortality likelihood of an individual patient presenting to the ED: the RISKINDEX. Machine learning models were able to discriminate between patients who died or survived within 31-days as depicted by AUROCs of 0.944 [0.935-0.951], 0.978 [0.973-0.982], 0.877 [0.866-0.888] and 0.904 [0.894-0.914] for Maastricht, Amersfoort, Sittard and Heerlen, respectively (Figure 1A). After calibration of the raw output scores (see supplemental Figure 1), the RISKINDEX corresponded well with the likelihood of 31-day mortality (Figure 1B) confirmed by brier scores of 0.033 [0.031-0.036], 0.015 [0.014 – 0.017], 0.044 [0.041-0.047] and 0.034 [0.032-0.036] for Maastricht, Amersfoort, Sittard and Heerlen, respectively. Hence, the RISKINDEX provides an individualized and precise assessment of 31-day mortality risk by combining patient characteristics and all available laboratory data available within the first two hours after ED presentation.

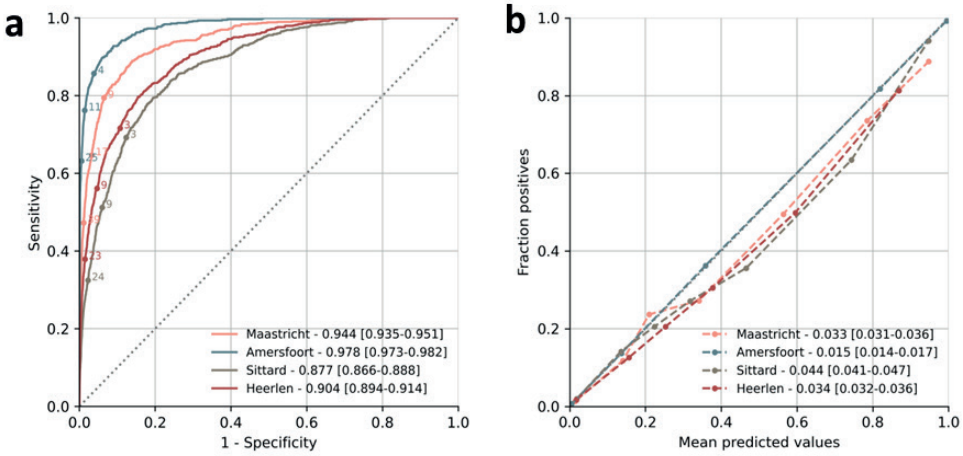


Figure 1. Discrimination and calibration of machine learning models. (A) Receiver operating characteristic curves (ROC) showing the discrimination of the LightGBM models in each of the different centers. Annotated points depict example RISKINDEX thresholds for illustrative purposes. (B) Calibration of the machine learning models with the observed proportion of 31-day mortality in each of the centers. Each point represents 10% of the patients in the validation dataset.

From theoretical model to clinical decision support tool

The RISKINDEX is by design a continuous measure, with a high RISKINDEX translating to a high likelihood of 31-day mortality and low RISKINDEX translating to low likelihood of 31-day mortality (see Supplementary Figures 1-4). We recognize however that in clinical practice most decision support tools use fixed thresholds to categorize patients as low-, medium- or high-risk. Consequently, our RISKINDEX can readily be transformed to such a fixed-threshold decision support tool, and users can control thresholds in accordance to the desired risk tolerance level. An illustrative example of how an individual hospital may employ the RISKINDEX is as follows: define the acceptable percentage of patients that are erroneously identified as "low-risk" by the model (any number from 0-100). This percentage, e.g. 1%, could be derived from an inventory of acceptable risk tolerance for adverse events by patients, health care workers, or both [47]. Then, use the corresponding negative predictive value (in this case 99%) to derive the matching RISKINDEX threshold from the calibration set and associated values for sensitivity, specificity, and proportion of subjects identified as low risk (Table 2). A similar approach can be applied to identify high-risk patients: define the positive predictive value that would provide an acceptable balance between true high risk patient identification and false positives, e.g. a positive predictive value of 75% would categorize between 1.1% and 3.9% as high-risk individuals with 1 in 4 "flaggings" by the clinical decision support tool being false positive (Table 2). A higher proportion of high-risk subject identification is feasible but will be at the expense of increased false positive flaggings.

Table 2. Illustrative example of clinical decision support tool using developed machine learning models. The fixed-threshold decision support tool was fixed at a negative predictive value of 99% to identify low-risk patients (corresponding RISK^{INDEX} cut-offs between 1.0 and 12.1) and fixed at a positive predictive value of 75% to identify high-risk patients (corresponding RISK^{INDEX} cut-offs between 29.9 and 64). Diagnostic metrics and proportion of patients identified as either low- or high-risk are described in the table.

Low-risk defined at a negative predictive value of 99%					
	RISK^{INDEX} cut-off	**Sensitivity**	**Specificity**	**NPV**	**Proportion of patients**
Maastricht	4.3	86.9% [84.2% - 88.9%]	87.7% [87.2% - 88.2%]	99.0% [98.8% - 99.2%]	83.0% [82.3% - 83.6%]
Meander	12.1	75.9% [72.3% - 78.6%]	97.5% [97.3% - 97.7%]	99.0% [98.8% - 99.1%]	94.7% [94.3% - 95.0%]
Sittard	1.0	85.5% [83.2% - 87.3%]	75.0% [74.2% - 75.7%]	98.8% [98.6% - 99.0%]	71.5% [70.7% - 72.0%]
Heerlen	1.0	82.7% [80.0% - 85.2%]	81.2% [80.8% - 81.8%]	98.9% [98.8% - 99.1%]	78.1% [77.7% - 78.8%]
High-risk defined at positive predictive value of 75%					
	RISK^{INDEX} cut-off	**Sensitivity**	**Specificity**	**PPV**	**Proportion of patients**
Maastricht	41.1	45.7% [42.3% - 49.5%]	99.0% [98.8% - 99.1%]	74.9% [70.8% - 77.8%]	3.9% [3.5% - 4.2%]
Meander	29.9	60.6% [56.9% - 64.1%]	99.2% [99.0% - 99.3%]	74.9% [70.8% - 78.3%]	3.1% [2.9% - 3.4%]
Sittard	64	14.7% [12.6% - 16.7%]	99.7% [99.6% - 99.8%]	74.8% [68.8% - 80.6%]	1.1% [1.0% - 1.3%]
Heerlen	41.1	27.1% [24.2% - 29.4%]	99.5% [99.4% - 99.7%]	74.8% [70.6% - 80.6%]	1.7% [1.5% - 2.0%]

Explainable model predictions

To understand the patterns underlying the RISK^{INDEX} generated for each individual patient, the SHAP algorithm was applied (see Supplemental Figures 6-8). This is illustrated for a low-, medium and high-risk individual in Figure 2. A high RISK^{INDEX} was generated for a 71-year old (+5.5 RISK^{INDEX}) individual with a high numeric amount of laboratory measurements (+31.5 RISK^{INDEX}), a high lactate dehydrogenase (LD; +17.7) and a low albumin level (+6), whilst the remainder of the features contributed another significant portion (+10.6) ultimately leading to a RISK^{INDEX} of 76. On the other hand, the low-risk individual had a normal albumin level (-0.9 RISK^{INDEX}), the presence of a pH blood measurement (-0.6), a normal lymphocyte level (-0.5) and the remaining variables further lowered the prediction (-1.3). Yet, a relatively high urea level (17 mmol/L) caused a small increase in RISK^{INDEX} (+0.7). These figures exemplify the explainability of our machine learning models.

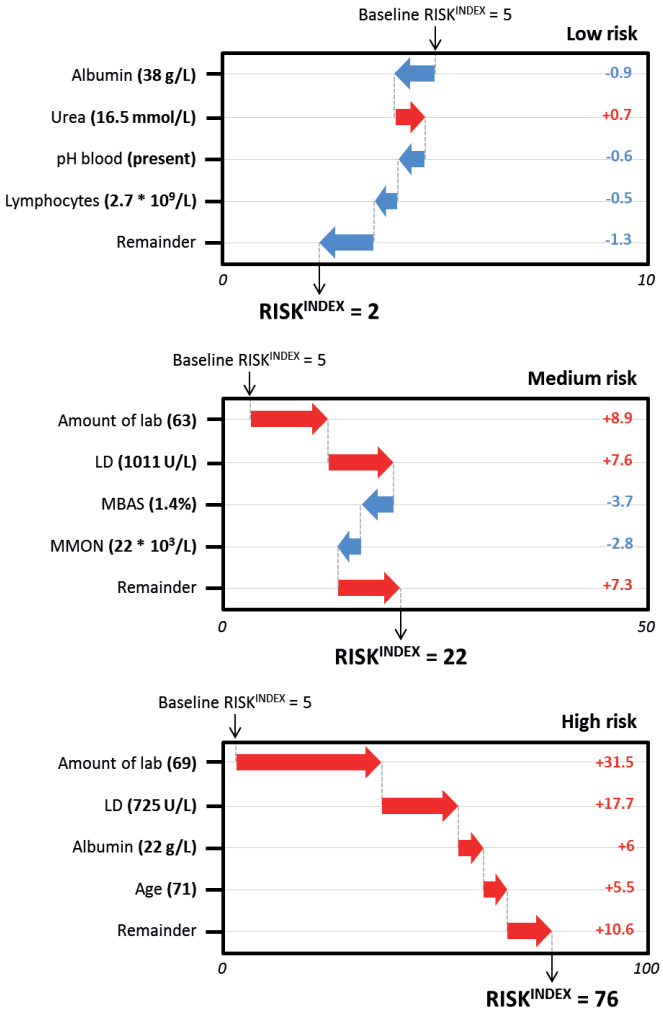


Figure 2. Illustrative example of how patient characteristics and laboratory results build up to a RISK^{INDEX}. Illustrative example of how patient characteristics and laboratory results build up to a RISK^{INDEX} in a low-risk (upper panel, RISK^{INDEX} of 2), medium-risk (mid panel, RISK^{INDEX} of 22), and high-risk patient (lower panel, RISK^{INDEX} of 76).

Discussion

In a large, multi-center study including over 260,000 patients presenting to the emergency department across four hospitals, machine learning technology was used to develop and evaluate novel clinical decision support tools that incorporate baseline patient characteristics and laboratory data to accurately predict the likelihood of 31-day mortality. Explainable machine learning models were utilized that were well calibrated and had overall high diagnostic performance. Our study has several unique characteristics.

First, our clinical decision support tool provides an individualized, precise, and rapid assessment of 31-day mortality risk -the RISKINDEX- by using patient characteristics and baseline laboratory results acquired within two hours after the presentation of the patient. Our models had high diagnostic performance with AUROCS of 0.944 [0.935-0.951], 0.978 [0.973-0.982], 0.877 [0.866-0.888] and 0.904 [0.894-0.914] (Maastricht, Amersfoort, Sittard and Heerlen), outperforming any clinical decision support tool or risk score currently used in the emergency department for risk stratification [6, 7, 48].

Second, the Shapley additive explanations (SHAP) algorithm was used to obtain explainable machine learning models [21, 23]. The SHAP algorithm facilitates the interpretation of patient characteristics and laboratory results that drive patient-specific RISKINDEX predictions (as illustrated in Figure 2). Development of such explainable machine learning models mitigates the issue of "black-box" predictions, and contributes to the understanding and acceptance of these models amongst clinicians and nurses. Furthermore, transparency in these models will likely become inevitable as international regulators have expressed concerns regarding black-box predictions, signaling that automated prediction systems are enforced to inform users about the logic involved, as well as the significance and the envisaged consequences of its predictions in the near future [49-51].

Third, our clinical decision support tool is highly versatile as it can be adjusted to the demands of each specific healthcare system or institution. For example, a triage algorithm was illustrated using a negative predictive value of 99% to identify low-risk patients, and a positive predictive value of 75% to identify high-risk patients (Table 2). Nevertheless, in case a more conservative policy would be desired, the low-risk thresholds can be adjusted accordingly, e.g. to an even higher NPV of 99.5% implying that only 5 out of a 1.000 patients would erroneously be identified as "low-risk". Implementation of such a triage system using our proposed clinical decision tool is convenient as current models rely on data that are easily collected through existing laboratory system infrastructure. This is an advantage compared to machine learning models trained with e.g. unstructured clinical data that require manual annotation or natural language processing.

Fourth, application of our methodology to four separate hospitals provides support for its

robustness and consistency. Differences in diagnostic performance between hospitals can in part be explained by demographic differences, patient mix (and hence baseline mortality rates), and the laboratory testing patterns of the attending physicians.

Last, the large sample size of more than 260,000 patients and 7.1 million laboratory tests allowed for the development of machine learning models with high performance. Despite our models being trained almost exclusively with laboratory data, they outperform machine learning models which also had full access to clinical data of a patient [18, 19]. This highlights that high-performance machine learning models -besides having access to as much individual patient data- require large sample size in order to achieve optimal performance for a specific prediction task.

Literature

A limited number of attempts to use machine learning technology for risk stratification in the ED in a retrospective setting have been described [12-14, 18, 19]. Klug et al. and Perng et al. developed machine learning models with performance in line with our study (AUCs of 0.96 and 0.93, respectively) [18, 19]. Although diagnostic performance was similar, there are some notable differences. First, these studies focused on populations from a single center. Second, these studies used clinical and vital characteristics of patients whereas our study almost exclusively relied on the laboratory results, which makes it easier to implement and extend to other hospitals. Third, the interpretation of our generated RISKINDEX was facilitated on a patient level using the recent SHAP algorithm. Fourth, illustrative implementation strategies were provided using pre-defined safety (NPV) and efficacy (PPV) measures to identify low- and high-risk patients at the emergency department, respectively.

Limitations

Several limitations should be recognized. First, the current study is based on retrospective data and prospective studies are desired to study performance and true clinical benefit of our clinical decision support tool in a real-world setting. This would also allow us to study the (dis) advantages of implementing these models using a triage system based on statistical thresholds (e.g. Figure 3) compared to an approach based on individual RISKINDEX estimates. Second, these models possess -despite being explainable- algorithmic bias; models have been trained entirely upon the basis of what humans have done before. This implies that the predictions of our model cannot be extrapolated, and that predictions in e.g. minority populations have a higher degree of uncertainty. To facilitate the interpretation of such uncertain predictions, it would be interesting to implement uncertainty measures amongst the prediction, e.g. in form of confidence intervals. This could warn clinicians when a certain prediction is highly uncertain, potentially leading to increased trust and interpretability amongst the users of these clinical decision support tools. This is novel development in machine learning which requires additional technical developments.

Conclusion

Our novel RISKINDEX clinical decision support tool incorporates patient characteristics and laboratory tests available within the first two hours after presentation to provide an individual, precise and rapid assessment of the patient's mortality risk within 31 days. These models had overall high diagnostic performance, are explainable, and can be implemented in a triage system extending current systems used in modern emergency departments. Prospective, follow-up studies are warranted to study the performance and clinical benefit of these models in a real-world clinical setting.

References

1. LaCalle, E. and E. Rabin, Frequent users of emergency departments: the myths, the data, and the policy implications. Ann Emerg Med, 2010. 56(1): p. 42-8.
2. Hooker, E.A., P.J. Mallow, and M.M. Oglesby, Characteristics and Trends of Emergency Department Visits in the United States (2010-2014). J Emerg Med, 2019. 56(3): p. 344-351.
3. Wansink, L., et al., Trend analysis of emergency department malpractice claims in the Netherlands: a retrospective cohort analysis. Eur J Emerg Med, 2019. 26(5): p. 350-355.
4. Guttmann, A., et al., Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ, 2011. 342: p. d2983.
5. Seymour, C.W., et al., Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 2016. 315(8): p. 762-74.
6. Olsson, T., A. Terent, and L. Lind, Rapid Emergency Medicine Score can predict long-term mortality in nonsurgical emergency department patients. Acad Emerg Med, 2004. 11(10): p. 1008-13.
7. Vorwerk, C., et al., Prediction of mortality in adult emergency department patients with sepsis. Emerg Med J, 2009. 26(4): p. 254-8.
8. Christ, M., et al., Modern triage in the emergency department. Dtsch Arztebl Int, 2010. 107(50): p. 892-8.
9. Crowe, C.A., et al., Comparison of severity of illness scoring systems in the prediction of hospital mortality in severe sepsis and septic shock. J Emerg Trauma Shock, 2010. 3(4): p. 342-7.
10. Collins, G.S., et al., External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol, 2014. 14: p. 40.
11. Ha, D.T., et al., Prognostic performance of the Rapid Emergency Medicine Score (REMS) and Worthing Physiological Scoring system (WPS) in emergency department. Int J Emerg Med, 2015. 8: p. 18.
12. Taylor, R.A., et al., Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med, 2016. 23(3): p. 269-78.
13. Shafaf, N. and H. Malek, Applications of Machine Learning Approaches in Emergency Medicine; a Review Article. Arch Acad Emerg Med, 2019. 7(1): p. 34.
14. Tang, F., et al., Predictive modeling in urgent care: a comparative study of machine learning approaches. JAMIA Open, 2018. 1(1): p. 87-98.
15. Levin, S., et al., Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. Ann Emerg Med, 2018. 71(5): p. 565-574 e2.
16. Peck, J.S., et al., Generalizability of a simple approach for predicting hospital admission from an emergency department. Acad Emerg Med, 2013. 20(11): p. 1156-63.
17. Barnes, S., et al., Real-time prediction of inpatient length of stay for discharge prioritization. J Am Med Inform Assoc, 2016. 23(e1): p. e2-e10.
18. Klug, M., et al., A Gradient Boosting Machine Learning Model for Predicting Early Mortality in the Emergency Department Triage: Devising a Nine-Point Triage Score. J Gen Intern Med, 2020. 35(1): p. 220-227.
19. Perng, J.W., et al., Mortality Prediction of Septic Patients in the Emergency Department Based on Machine Learning. J Clin Med, 2019. 8(11).
20. van Doorn, W.P.T.M., et al., A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. medRxiv, 2020: p. 2020.11.24.20237636.
21. Lundberg, S.M., et al., From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020.
22. Lundberg, S.M., et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng, 2018. 2(10): p. 749-760.
23. Thorsen-Meyer, H.-C., et al., Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital

Health, 2020.
24. Hyland, S.L., et al., Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med, 2020. 26(3): p. 364-373.
25. von Elm, E., et al., The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet, 2007. 370(9596): p. 1453-7.
26. World Medical, A., World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA, 2013. 310(20): p. 2191-4.
27. Bagley, S.C., H. White, and B.A. Golomb, Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol, 2001. 54(10): p. 979-85.
28. Harrell, F.E., Jr., et al., Regression models for prognostic prediction: advantages, problems, and suggested solutions. Cancer Treat Rep, 1985. 69(10): p. 1071-77.
29. Forsstrom, J.J. and K.J. Dalton, Artificial neural networks for decision support in clinical medicine. Ann Med, 1995. 27(5): p. 509-17.
30. Agatonovic-Kustrin, S. and R. Beresford, Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. J Pharm Biomed Anal, 2000. 22(5): p. 717-27.
31. Zhang, Z., et al., Predictive analytics with gradient boosting in clinical medicine. Ann Transl Med, 2019. 7(7): p. 152.
32. Chen, T. and C. Guestrin XGBoost: A Scalable Tree Boosting System. arXiv e-prints, 2016.
33. Ke, G., et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017: p. 3146--3154.
34. Prokhorenkova, L., et al. CatBoost: unbiased boosting with categorical features. arXiv e-prints, 2017.
35. Podgorelec, V., et al., Decision trees: an overview and their use in medicine. J Med Syst, 2002. 26(5): p. 445-63.
36. Bergstra, J., et al., Algorithms for hyper-parameter optimization. Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011, Granada, Spain: Curran Associates Inc. 2546–2554.
37. Guo, C., et al. On Calibration of Modern Neural Networks. arXiv e-prints, 2017.
38. Advances in Large Margin Classifiers, ed. J.S. Alexander and J.B. Peter. 2000: MIT Press. 412.
39. Zadrozny, B. and C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002, ACM: Edmonton, Alberta, Canada. p. 694-699.
40. Kumar, A., P. Liang, and T. Ma Verified Uncertainty Calibration. arXiv e-prints, 2019.
41. BRIER, G.W., VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. Monthly Weather Review, 1950. 78(1): p. 1-3.
42. Niculescu-Mizil, A. and R. Caruana, Predicting good probabilities with supervised learning, in Proceedings of the 22nd international conference on Machine learning. 2005, ACM: Bonn, Germany. p. 625-632.
43. Molnar, C., Interpretable Machine Learning, in A Guide for Making Black Box Models Explainable. 2020
44. Lipovetsky, S. and M. Conklin, Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry, 2001. 17(4): p. 319-330.
45. Štrumbelj, E. and I. Kononenko, Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems, 2013. 41: p. 647-665.
46. The Shapley Value: Essays in Honor of Lloyd S. Shapley. 1988, Cambridge: Cambridge University Press.
47. Brown, T.B., et al., Assessment of risk tolerance for adverse events in emergency department chest pain patients: a pilot study. J Emerg Med, 2010. 39(2): p. 247-52.
48. Chang, S.H., et al., Performance Assessment of the Mortality in Emergency Department Sepsis Score, Modified Early Warning Score, Rapid Emergency Medicine Score, and Rapid Acute Physiology Score in Predicting Survival Outcomes of Adult Renal Abscess Patients in the Emergency Department. Biomed Res Int, 2018. 2018: p. 6983568.
49. Jobin, A., M. Ienca, and E. Vayena, The global landscape of AI ethics guidelines. Nature Machine Intelligence, 2019.
50. Cohen, I.G., et al., The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood), 2014. 33(7): p. 1139-47.
51. EU, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). 2016, Off J Eur Communities. p. 1-88.

Supplemental material

Supporting information

Supplemental section A. Information of baseline model architectures.

A comparison of available algorithms was conducted on the 31-day mortality prediction task. The following algorithms:

- Logistic regression: a simple logistic regression model was used with L2 regularization, a tolerance of 1e-4 and a maximum amount of iterations of 1,000. We used the limited Broyden–Fletcher–Goldfarb–Shanno (lbfgs) algorithm as optimizer. Logistic regression was implemented using Python (version 3.7.1) and the scikit-learn (version 0.22.1) package.
- Feed-forward neural networks: a simple feed-forward, multi-layer perceptron neural network was implemented consisting of three hidden layers with respectively 128, 64 and 32 neurons and RELU activation functions. We trained the network using a constant learning rate of 0.001 with a batch size of 1 and the Adam optimization scheme. The neural network was implemented using Python programming language (version 3.7.1) using packages Keras (version 2.2.2) and scikit-learn (version 0.22.1).
- Random Forest: a random forest classifier with a decision tree as base learner, consisting of 300 trees with gini criterion and a maximum depth of 50 was used. We used bootstrapped samples for building trees. Random forest was implemented using Python (version 3.7.1) and the scikit-learn (version 0.22.1) package
- Gradient-boosting systems: we used three different implementations of gradient-boosting systems: CatBoost [1], XGBoost [2] and LightGBM [3]. Each of them has specific unique implementation details, but all use decision trees as the base weak learner and gradient boosting to iteratively fit a sequence of such trees. To provide a valuable baseline comparison, we aimed to evaluate these implementations with as identical hyperparameters as possible. Thus, for each implementation, we used a learning rate of 0.01, a maximum number of trees of 500 and a maximum depth of each base learner to be 50. We implemented this using the Python programming language (version 3.7.1) using packages XGBoost (version 0.90), CatBoost (version 0.20.2) and LightGBM (version 2.3.1).

Supporting Tables

Supplemental Table 1. Baseline comparison of statistical and machine learning models.
Comparison of statistical and machine learning models for the prediction of 31-day mortality in validation dataset. Models were trained on the training dataset and their performance was evaluated on the tuning dataset. Performance was assessed using their discrimination ability by ROC curves, and their calibration performance by using Brier scores. 95% confidence intervals were calculated using 1,000 bootstraps.

Baseline model	Maastricht		Sittard		Heerlen		Amersfoort	
	AUC	Brier	AUC	Brier	AUC	Brier	AUC	Brier
Logistic regression	0.73 [0.708 – 0.748]	0.06 [0.060 – 0.068]	0.62 [0.581 – 0.625]	0.06 [0.055 – 0.066]	0.67 [0.655 – 0.694]	0.05 [0.049 – 0.054]	0.79 [0.767 – 0.803]	0.04 [0.034 – 0.039]
Neural network	0.76 [0.740 – 0.779]	0.06 [0.055 – 0.063]	0.78 [0.751 – 0.812]	0.06 [0.048 – 0.059]	0.78 [0.759 – 0.796]	0.04 [0.039 – 0.045]	0.81 [0.788 – 0.823]	0.04 [0.038 – 0.043]
Random Forest	0.77 [0.752 – 0.789]	0.05 [0.051 – 0.057]	0.76 [0.743 – 0.777]	0.05 [0.050 – 0.052]	0.76 [0.738 – 0.773]	0.04 [0.040 -0.045]	0.80 [0.781 – 0.817]	0.04 [0.032 – 0.041]
CatBoost	0.89 [0.879 – 0.901]	0.05 [0.042 – 0.047]	0.86 [0.855 – 0.0873]	0.05 [0.042 – 0.049]	0.90 [0.898 – 0.909]	0.03 [0.030 – 0.037]	**0.94 [0.924 - 0.951]**	**0.02 [0.016 – 0.02]**
XGBoost	0.87 [0.856 – 0.880]	0.05 [0.045 – 0.051]	0.87 [0.857 – 0.878]	0.05 [0.043 – 0.049]	0.87 [0.865 – 0.890]	0.04 [0.034 – 0.040]	0.92 [0.913 – 0.931]	0.03 [0.024 – 0.035]
LightGBM	**0.91 [0.898 – 0.912]**	**0.04 [0.037 – 0.042]**	**0.88 [0.869 – 0.889]**	**0.04 [0.041 – 0.047]**	**0.90 [0.897 – 0.911]**	**0.03 [0.031 – 0.036]**	0.93 [0.922 – 0.948]	0.02 [0.019 – 0.028]

Supplemental Table 2. Hyperparameter search space.
Hyperparameter combinations evaluated in the current study to find optimal hyperparameters for LightGBM architectures.

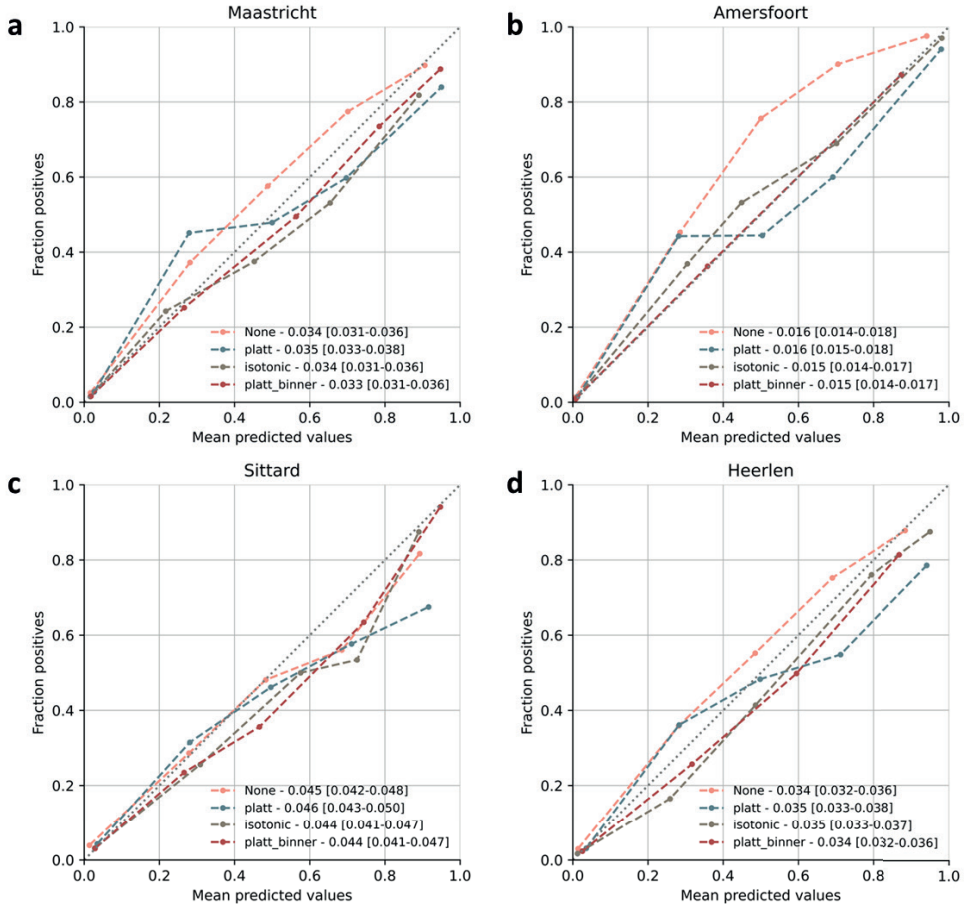
Hyperparameter	Values considered
Data preprocessing	
Normalization to [0, 1]	On, off
Presence of binary 'presence' variable	On, off
Presence of numeric amount variable	On, off
Gradient-boosting tree architecture	
Number of boosted trees to fit	10 – 1000 on a 10 scale
Maximum amount of tree leaves for base learner	10 – 1000 on a 10 scale
Maximum tree depth for base learner	1 – 50 on a 1.0 scale
Minimum sum of instance weight needed in a leaf	0 – 10 on a 1.0 scale
Subsample ratio of columns when constructing each tree.	0.01 to 1.0 on a 0.01 scale
Model training	
Learning rate	0.025 – 1.0 on a 0.005 scale
Learning rate scheduling	On, off
Learning rate scheduling decay	0.5, 0.6, 0.7, 0.8, 0.9, 0.95
Subsample ratio of the training instance.	0.5 to 1.0 on a 0.05 scale
L1 regularization term on weights.	1^{e}-9 to 1^{e}3 on a log-scale
L2 regularization term on weights.	1^{e}-9 to 1^{e}3 on a log-scale

Supplemental Table 3. Hyperparameter settings for LightGBM models. Hyperparameter settings for each of the LightGBM models developed in Maastricht, Sittard, Heerlen and Amersfoort, respectively.

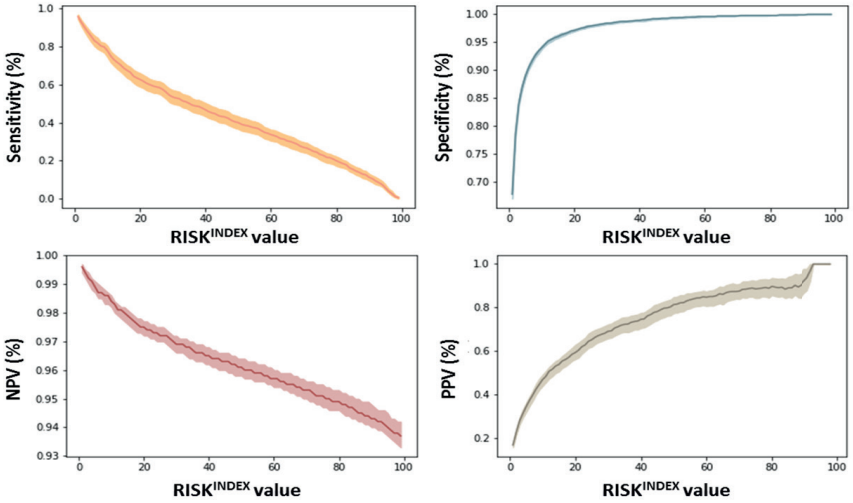
Hyperparameter	MUMC Maastricht	Zuyderland Sittard	Zuyderland Heerlen	Meander Amersfoort
Data preprocessing				
Normalization to [0, 1]	Off	Off	Off	Off
Presence of binary 'presence' variable	On	On	On	On
Presence of numeric amount variable	On	On	On	On
Gradient-boosting tree architecture				
Number of boosted trees to fit	370	690	310	420
Maximum amount of tree leaves for base learner	240	940	640	220
Maximum tree depth for base learner	23	14	13	47
Minimum sum of instance weight needed in a leaf	2.0	7.0	3.0	4.0
Subsample ratio of columns when constructing each tree	0.809	0.681	0.971	0.824
Model training				
Learning rate	0.075	0.090	0.135	0.145
Learning rate scheduling	On	On	On	On
Learning rate scheduling decay	0.8	0.8	0.7	0.8
Subsample ratio of the training instance.	0.75	0.70	0.75	0.65
L1 regularization term on weights.	2.435 e-9	4.512 e-7	3.743 e-7	2.799 e-9
L2 regularization term on weights.	4.593 e-6	5.329 e-1	4.529 e-2	1.069 e-6

Supplemental Figures

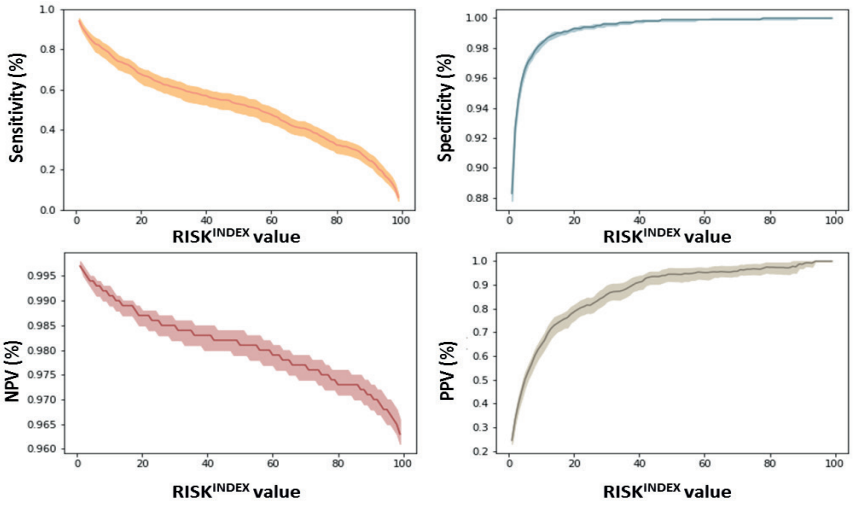
Supplemental Figure 1. Model calibration.
To ensure that we can make probabilistic interpretations of our proposed RISKINDEX, we need to confirm that the predicted RISKINDEX corresponds with the likelihood of 31-day mortality. As expected due to the nature of our machine learning architecture, our raw, i.e. uncalibrated, models did not provide us with well-calibrated values (see A-D). Calibration with the Platt-Binner technique, calibration curves show well calibrated algorithms confirmed by low Brier scores of 0.033 [0.031-0.036], 0.015 [0.014-0.017], 0.044 [0.041-0.047] and 0.034 [0.032-0.036] for each of the centers as previously described.



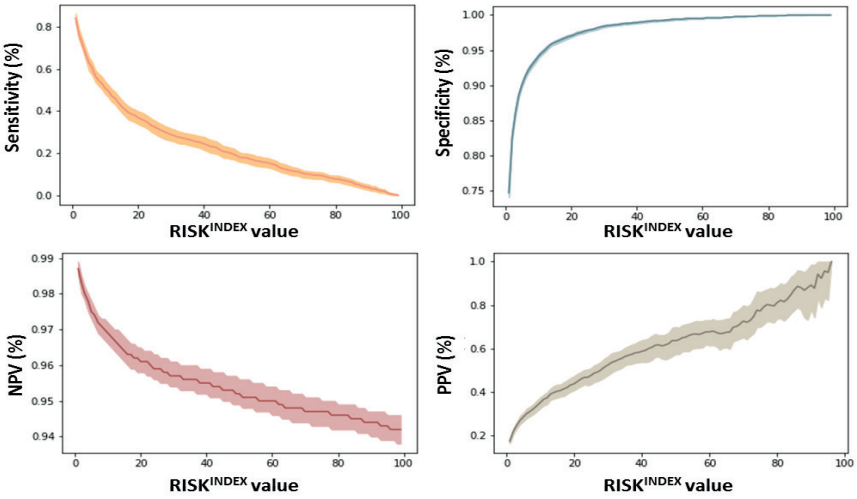
Supplemental Figure 2. Embedded reference figures for machine learning model in Maastricht.
For each patient in the Maastricht University Medical Center the model outputs a RISK^{INDEX} value which corresponds to an estimated sensitivity, specificity, PPV and NPV with 95% confidence intervals. Lines are the point estimates at each RISK^{INDEX} and shaded regions are the 95% confidence intervals.



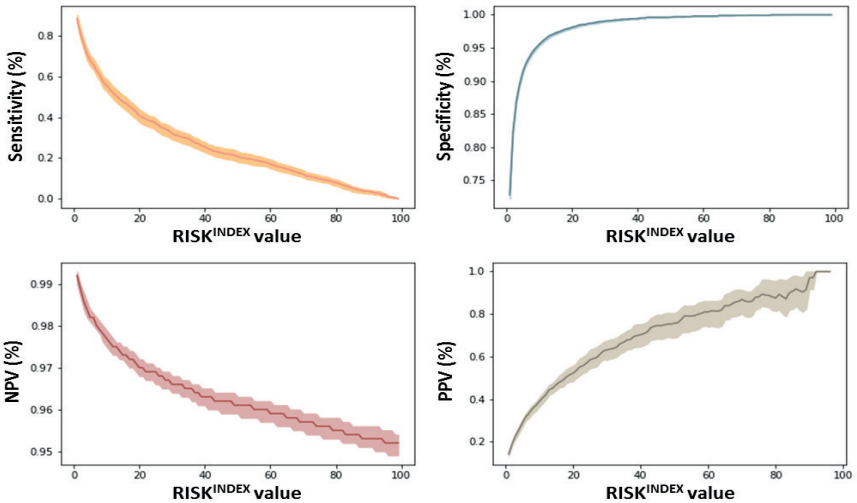
Supplemental Figure 3. Embedded reference figures for machine learning model in Amersfoort.
For each patient in the Meander Medical Center (Amersfoort) the model outputs a RISK^{INDEX} value which corresponds to an estimated sensitivity, specificity, PPV and NPV with 95% confidence intervals. Lines are the point estimates at each RISK^{INDEX} and shaded regions are the 95% confidence intervals.



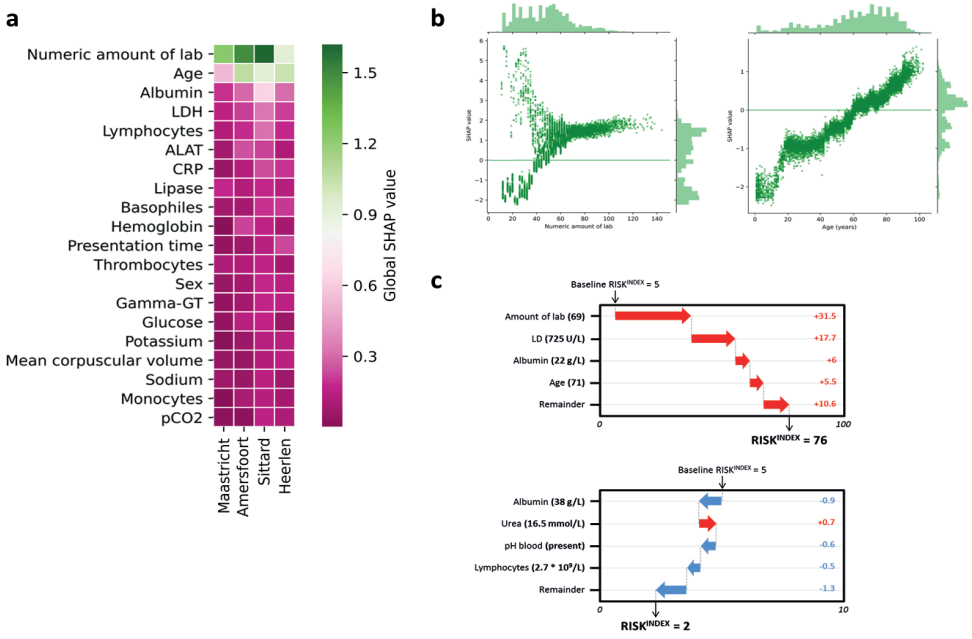
Supplemental Figure 4. Embedded reference figures for machine learning model in Sittard.
For each patient in the Zuyderland medical center, location Sittard the model outputs a RISK^{INDEX} value which corresponds to an estimated sensitivity, specificity, PPV and NPV with 95% confidence intervals. Lines are the point estimates at each RISK^{INDEX} and shaded regions are the 95% confidence intervals.



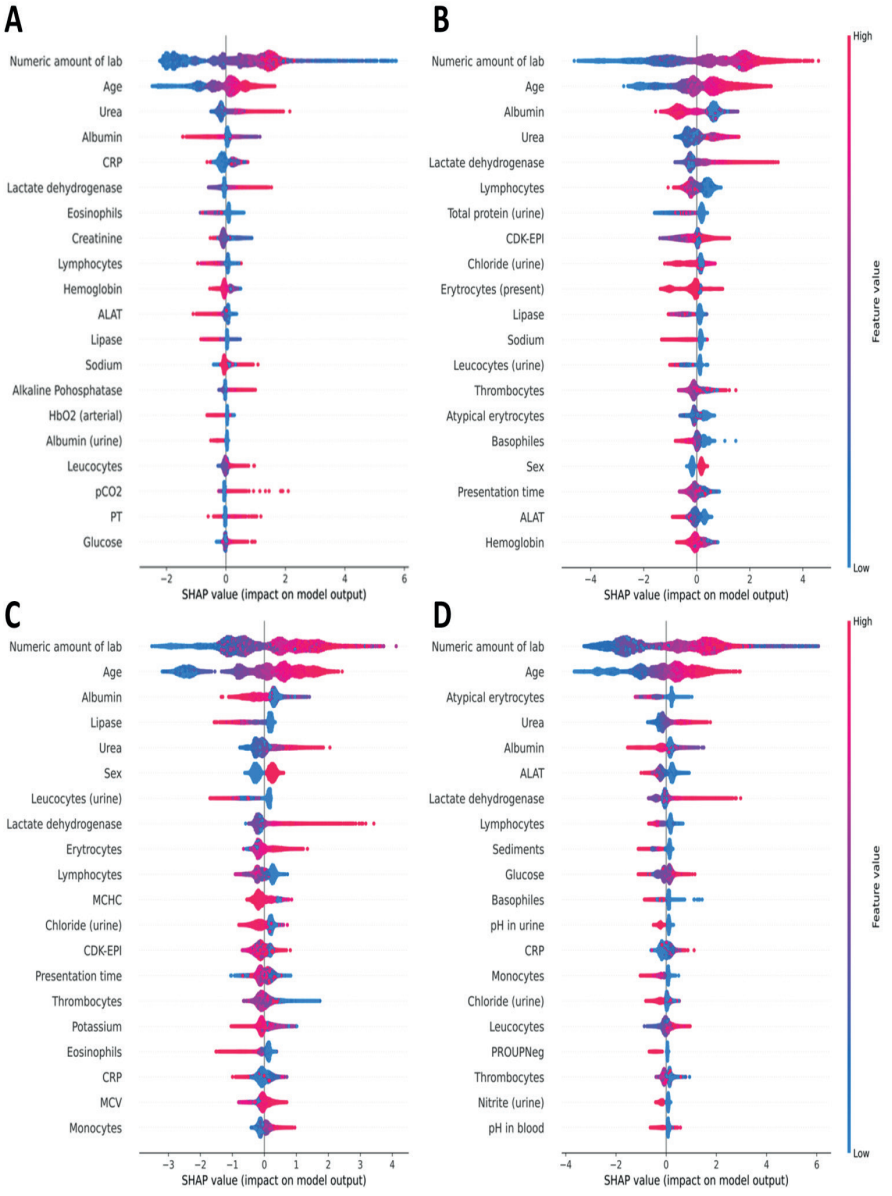
Supplemental Figure 5. Embedded reference figures for machine learning model in Heerlen**.**
For each patient in the Zuyderland medical center, location Heerlen the model outputs a RISK^{INDEX} value which corresponds to an estimated sensitivity, specificity, PPV and NPV with 95% confidence intervals. Lines are the point estimates at each RISK^{INDEX} and shaded regions are the 95% confidence intervals.

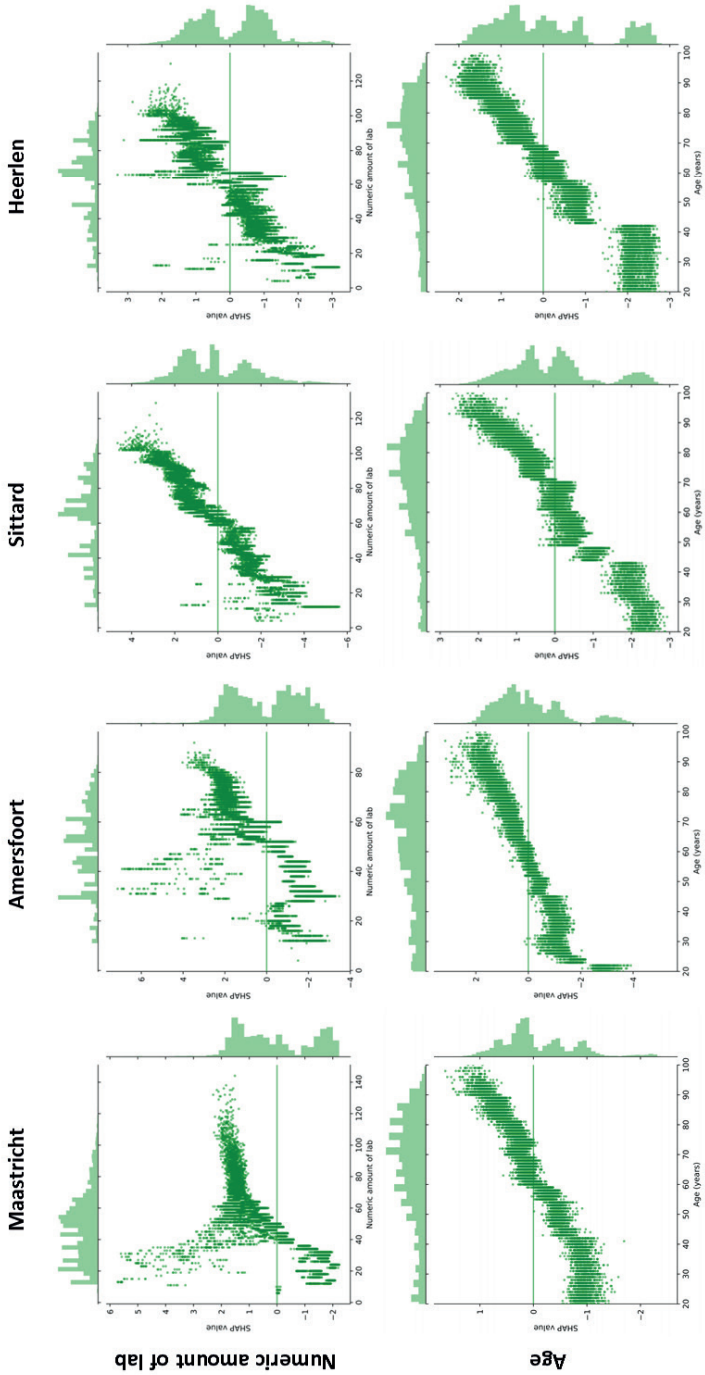


Supplemental Figure 6. Impact of model parameters.
(A) Heatmap of the global SHAP values for the developed models in each of the centers. Features were ordered from highest to lowest impact averaged over the four models. Green features indicate high impact; purple features indicate lower overall impact. (B) Specific examination of the relationship between the top-2 features (numeric amount of laboratory requests and age) and the SHAP value in the machine learning model of Maastricht. We found that the number of laboratory measurements possesses a 'double' relationship with the SHAP value, whilst the age of an individual shows a positive, almost linear relationship with the SHAP value. (C) Illustrative example of how patient characteristics and laboratory results build up to a $RISK^{INDEX}$ for a high- (upper panel) and low-risk (lower panel) individual.



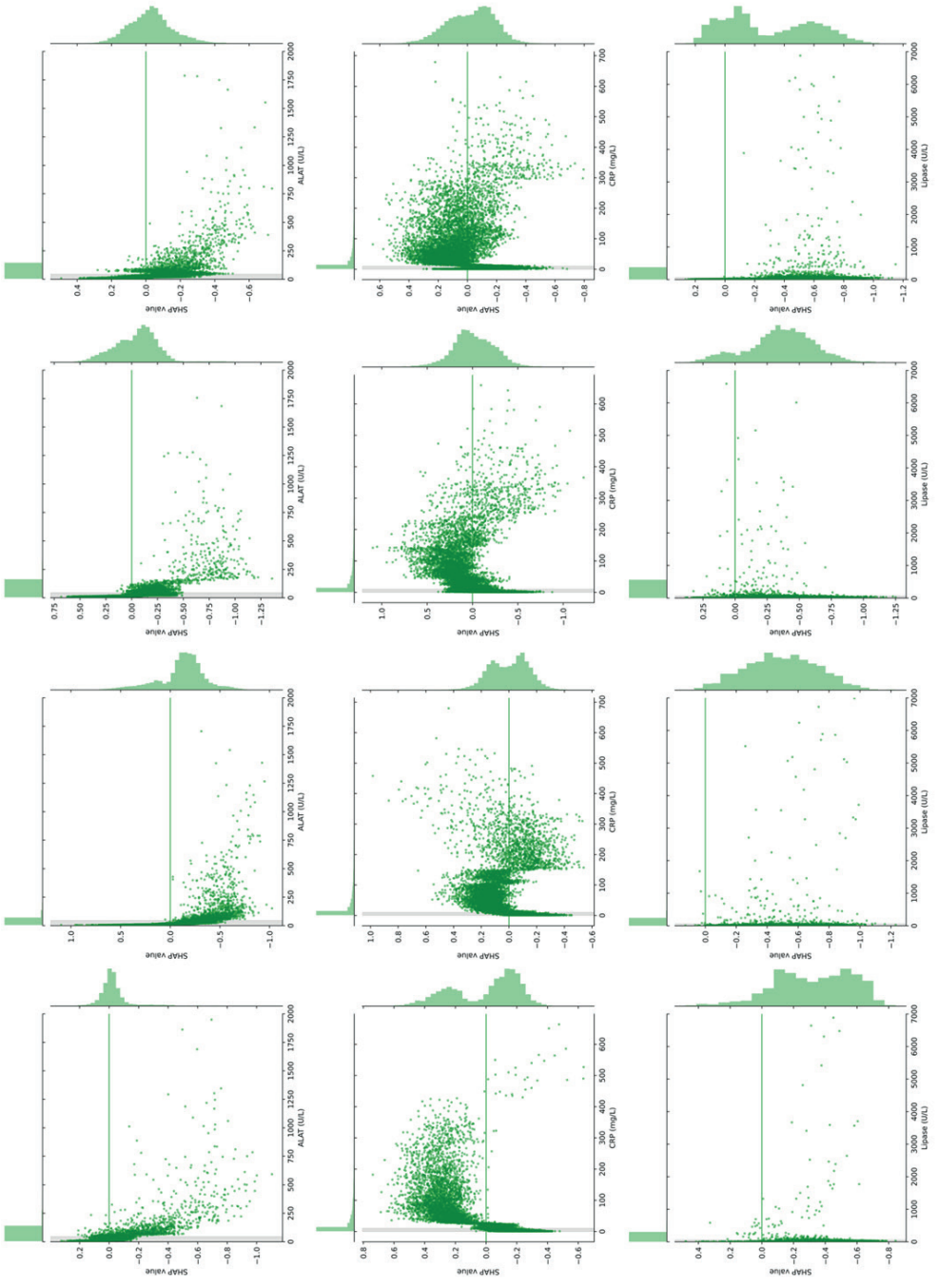
Supplemental Figure 7. Global SHAP values in each of the centers.
Analysis of feature importance in the LightGBM model developed in Maastricht (A), Sittard (B), Heerlen (C) and Meander (D) using SHAP values. The features are ranked by importance in descending order based on the sum of the SHAP values over all the samples.





Supplemental Figure 8. Impact of individual features on the SHAP value.
The relationship of each feature from the top-20 features with the SHAP value was evaluated for each of the centers (left to right: Maastricht, Amersfoort, Sittard and Heerlen). Individual points are depicted on the X-axis with their associated SHAP value on the Y-axis. Reference ranges -if available- for the specific laboratory test are shown as a grey area. Distributions of the values as well as the SHAP values are plotted on the outside borders of each graph.

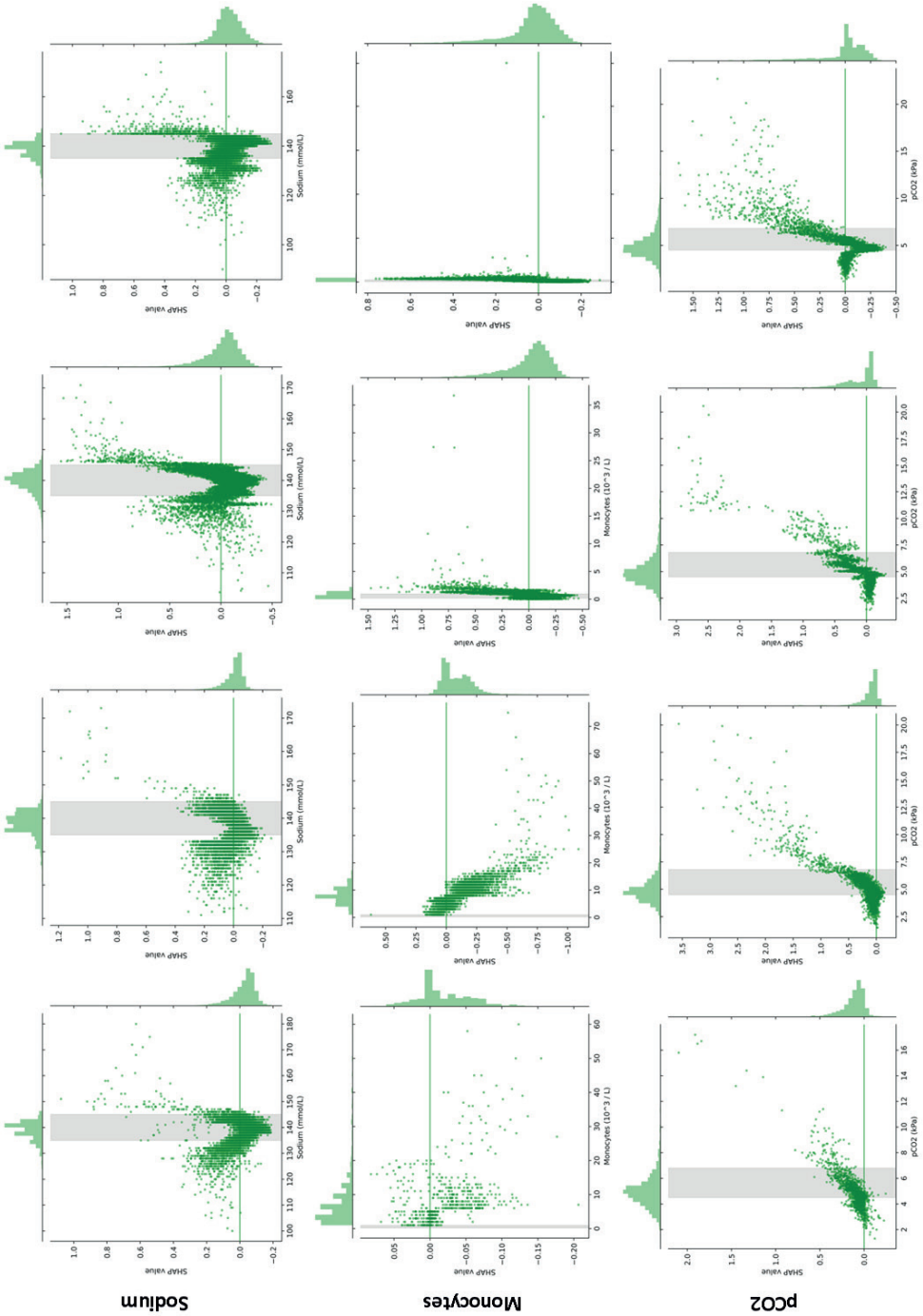
Albumin
LDH
Lymphocytes
SHAP value
Albumin (g/L)
LDH (U/L)
Lymphocytes (10^9 / L)


ALAT
CRP
Lipase
SHAP value
ALAT (U/L)
CRP (mg/L)
Lipase (U/L)

Basophiles
Hemoglobin
Presentation time
Basophiles (%)
Hemoglobin (mmol/L)
Presentation time (normalized)
SHAP value

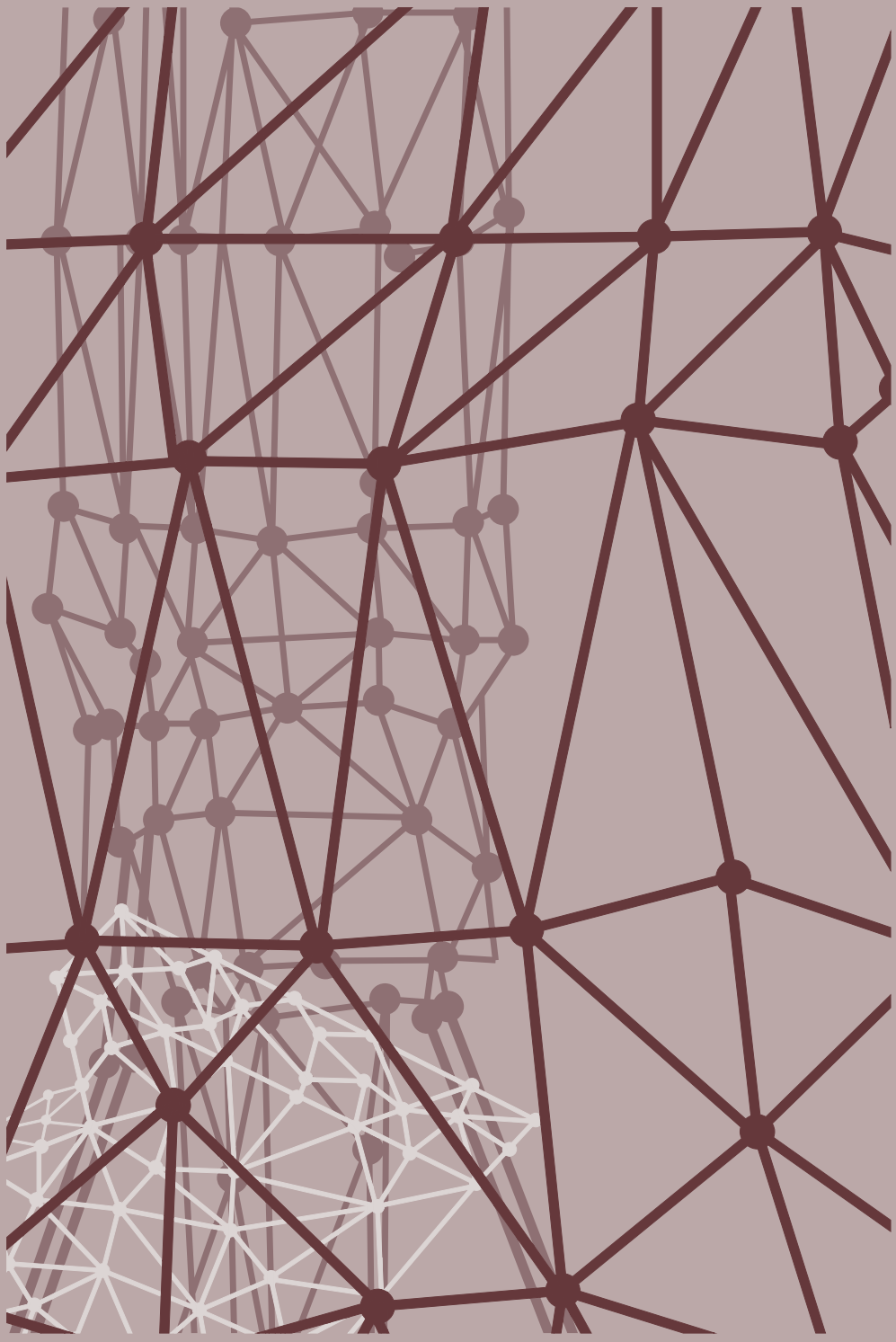
Thrombocytes
Sex (male)
Gamma-GT
SHAP value
Thrombocytes (10^9 / L)
Sex (male)
Gamma-GT (U/L)

Glucose
Potassium
Mean corpuscular volume
SHAP value
Glucose (mmol/L)
Potassium (mmol/L)
Mean corpuscular volume (fL)


pCO2
Monocytes
Sodium
SHAP value
pCO2 (kPa)
Monocytes (10^3 / L)
Sodium (mmol/L)

Supplemental references

1. Prokhorenkova, L., et al. CatBoost: unbiased boosting with categorical features. arXiv e-prints, 2017.
2. Chen, T. and C. Guestrin XGBoost: A Scalable Tree Boosting System. arXiv e-prints, 2016.
3. Ke, G., et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017: p. 3146--3154.



CHAPTER 9

MACHINE LEARNING FOR RISK STRATIFICATION IN PATIENTS WITH COVID-19 IN THE EMERGENCY DEPARTMENT

Paul M.E.L. van Dam, William P.T.M. van Doorn,
Steven J.R. Meex, Patricia M. Stassen

IN PREPARATION

Letter to the editor

To lessen the burden of the Coronavirus disease 2019 (COVID-19) pandemic, it is essential to rapidly and adequately identify emergency department (ED) patients' probability of poor or good outcomes.[1,2] Although most patients with COVID-19 develop only mild symptoms, some develop severe and potentially fatal complications.[3,4] Recently, we described in this Journal an accurate machine learning (ML) model that is able to predict 31-day mortality in emergency department (ED) patients based on laboratory tests in the first two hours of the ED visit (the RISK^{INDEX} score), yielding an AUC of 0.94. The ML model was developed in a general population of ED patients. As a part of our recent retrospective cohort study to validate several prediction models in ED patients with COVID-19, we also used our ML model to calculate the RISK^{INDEX} score. The aim of the present substudy was to evaluate the discriminatory performance of the ML model with regard to poor outcome in ED patients with COVID-19.

We included 403 consecutive adult patients diagnosed with COVID-19 at the ED of the Maastricht University Medical Center+ (MUMC+), a combined secondary/tertiary care center in the Netherlands. Patients were included if they had symptoms compatible with COVID-19 and a positive result of the polymerase chain reaction (PCR) for SARS-CoV-2 in respiratory specimens or (very) high suspicion of COVID-19 according to the chest computed tomography (CT) scan (CORADS 4 or 5).[5] The medical ethics committee of the MUMC+ approved this study (METC 2020-1572).

From medical records, we collected data on age, sex, comorbidity, laboratory test results, admission to the medium care unit (MCU) or intensive care unit (ICU) and 30-day mortality. Data on mortality were verified using the medical records, which are connected to the municipal administration office. The ML model was used to calculate the RISK^{INDEX} score The primary outcome was all-cause mortality within 30 days of ED presentation. Secondary outcomes were all-cause mortality within 14 days, and a composite outcome of 30-day mortality and/or admission to the MCU/ICU. We determined the discriminatory performance of the ML model by calculation of the area under the receiver operating characteristics curve (AUC) with 95% confidence intervals (CI).

For all 403 ED patients with COVID-19 who were included during the study period follow up was complete. The median age of patients was 71 years (interquartile range 60-78) and 66.0% were male. In our sample, 95 patients died during follow up, yielding a 30-day mortality of 23.6% and a 14-day mortality of 19.1%. Sixty-six patients (16.4%) were admitted to ICU, 48 patients (11.9%) to MCU, and a total of 152 patients (37.7%) met the composite endpoint of 30-day mortality and/or admission to MCU/ICU. The RISK^{INDEX} score showed good discriminatory performance in this cohort with an AUC of 0.80 (95% CI: 0.76-

0.85) for 30-day mortality, an AUC of 0.78 (95% CI: 0.73-0.84) for 14-day mortality and an AUC of 0.76 (95% CI: 0.71-0.81) for the composite endpoint.

In this retrospective study, we externally validated the RISK^{INDEX} score for its ability to predict 30-day mortality, 14-day mortality or admission to MCU/ICU in ED patients with COVID-19. We found that the RISK^{INDEX} score had very good discriminatory performance. When comparing the RISK^{INDEX} score with the clinical prediction models we validated in our recent study, it is as accurate as the two scores with the highest discriminatory value in ED patients with COVID-19 (RISE UP score and 4C mortality score, which yielded AUCs of 0.83 and 0.84, respectively).

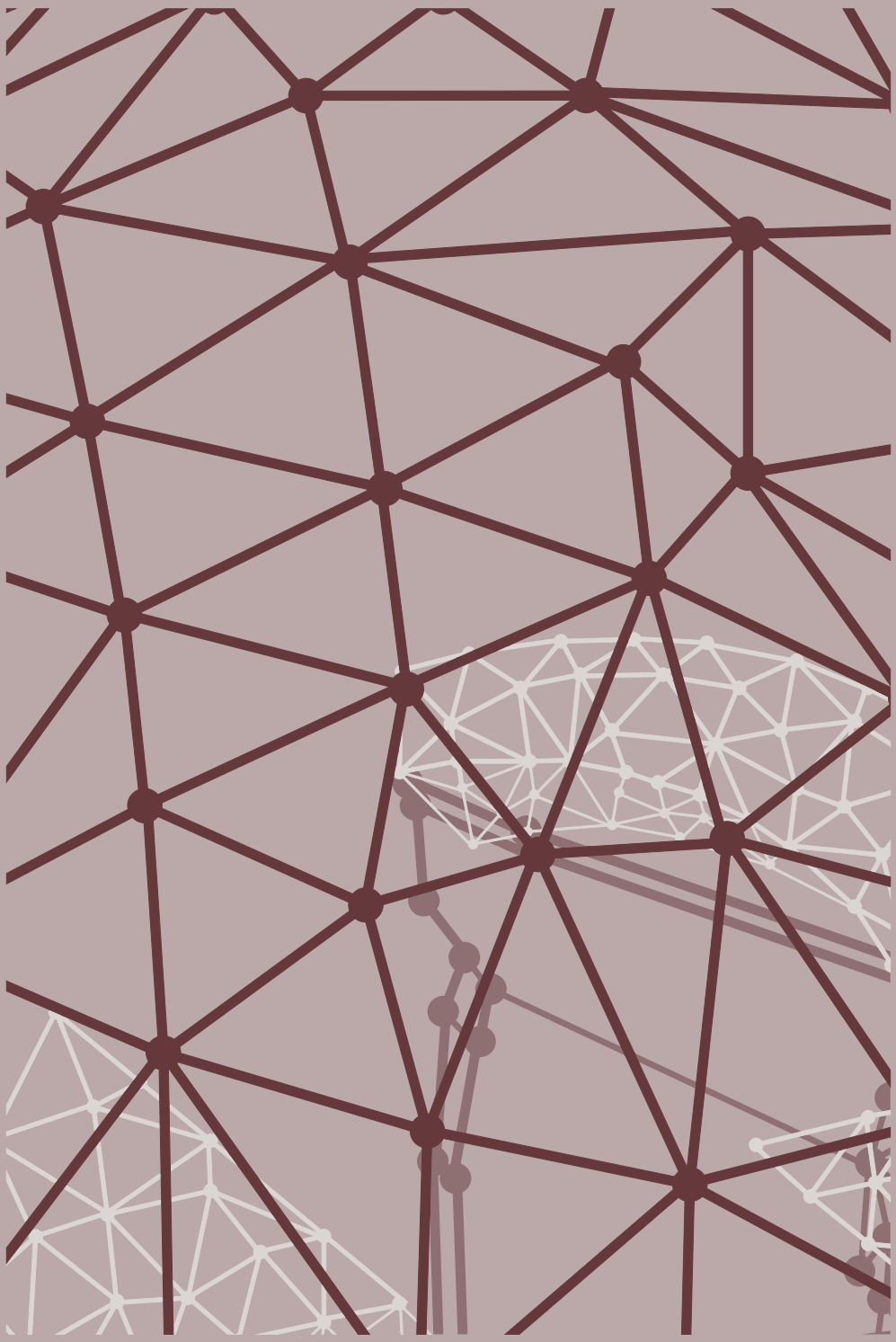
In previous studies the RISK^{INDEX} score showed excellent discriminatory performance with AUCs of 0.94. The 30-day mortality in our sample of ED patients with COVID-19 was 23.6%, which is much higher than the 31-day mortality in the general population of ED patients in previous studies (5.3%). This may explain the lower AUC in our sample, because the machine learning patterns are based on a sample with lower mortality rates. Furthermore, our study sample was much smaller than the sample in which the ML model was developed. Nevertheless, the RISK^{INDEX} score can be used to predict the probability of adverse outcome in the first two hours of the ED visit using only the results of routinely performed laboratory tests.

A drawback of our study was its single center design, which may limit the generalizability of the results. However, our cohort of patients with COVID-19 was relatively large and recruited in one of the most heavily affected areas of the Netherlands.

In conclusion, the RISK^{INDEX} score had high discriminatory performance for short term mortality and other adverse outcomes in ED patients with COVID-19. Therefore, the RISK^{INDEX} is useful to identify patients at high risk for poor outcome and may thus guide clinical decision-making. Further (prospective) research is needed to investigate to what extent clinical decision-making is influenced by the results of this ML model.

References

1. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507-13.
2. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg. 2020;76:71-6.
3. Guan WJ, Ni ZY, Hu Y, Liang WH, Ou CQ, He JX, et al. Clinical Characteristics of Coronavirus Disease 2019 in China. N Engl J Med. 2020;382(18):1708-20.
4. Bruggemann R, Gietema H, Jallah B, Ten Cate H, Stehouwer C, Spaetgens B. Arterial and venous thromboembolic disease in a patient with COVID-19: A case report. Thromb Res. 2020;191:153-5.
5. Prokop M, van Everdingen W, van Rees Vellinga T, Quarles van Ufford J, Stoger L, Beenen L, et al. CO-RADS - A categorical CT assessment scheme for patients with suspected COVID-19: definition and evaluation. Radiology. 2020:201473.



CHAPTER 10

MACHINE LEARNING-BASED GLUCOSE PREDICTION WITH USE OF CONTINUOUS GLUCOSE AND PHYSICALACTIVITY MONITORING DATA: THE MAASTRICHT STUDY

William P.T.M. van Doorn*, Yuri D. Foreman*, Nicolaas C. Schaper, Hans H.C.M. Savelberg, Annemarie Koster, Carla J.H. van der Kallen, Anke Wesselius, Miranda T. Schram, Ronald M.A. Henry, Pieter C. Dagnelie, Bastiaan E. de Galan, Otto Bekers, Coen D.A. Stehouwer, Steven J.R. Meex, Martijn C.G.J. Brouwers
* equal contribution

PLOS ONE.
2021 Jun 24;16(6):e0253125

Abstract

Introduction: Closed-loop insulin delivery systems, which integrate continuous glucose monitoring (CGM) and algorithms that continuously guide insulin dosing, have been shown to improve glycaemic control. The ability to predict future glucose values can further optimize such devices. In this study, we used machine learning to train models in predicting future glucose levels based on prior CGM and accelerometry data.

Methods: We used data from The Maastricht Study, an observational population-based cohort that comprises individuals with normal glucose metabolism, prediabetes, or type 2 diabetes. We included individuals who underwent >48h of CGM (n=851), most of whom (n=540) simultaneously wore an accelerometer to assess physical activity. A random subset of individuals was used to train models in predicting glucose levels at 15- and 60-minute intervals based on either CGM data or both CGM and accelerometer data. In the remaining individuals, model performance was evaluated with root-mean-square error (RMSE), Spearman's correlation coefficient (rho) and surveillance error grid. For a proof-of-concept translation, CGM-based prediction models were optimized and validated with the use of data from individuals with type 1 diabetes (OhioT1DM Dataset, n=6).

Results: Models trained with CGM data were able to accurately predict glucose values at 15 (RMSE: 0.19mmol/L; rho: 0.96) and 60 minutes (RMSE: 0.59mmol/L, rho: 0.72). Model performance was comparable in individuals with type 2 diabetes. Incorporation of accelerometer data only slightly improved prediction. The error grid results indicated that model predictions were clinically safe (15 min: >99%, 60 min >98%). Our prediction models translated well to individuals with type 1 diabetes, which is reflected by high accuracy (RMSEs for 15 and 60 minutes of 0.43 and 1.73 mmol/L, respectively) and clinical safety (15 min: >99%, 60 min: >91%).

Conclusions: Machine learning-based models are able to accurately and safely predict glucose values at 15- and 60-minute intervals based on CGM data only. Future research should further optimize the models for implementation in closed-loop insulin delivery systems.

Introduction

The increasing prevalence of diabetes entails an increase in debilitating complications, such as retinopathy, neuropathy, and cardiovascular disease[1-3]. Maintaining plasma glucose levels within the reference range is essential for the prevention of diabetes-related complications, which are generally attributable to chronic hyperglycaemia, although hypoglycaemia has been suggested to contribute to cardiovascular disease risk as well[3-5]. One of the most promising developments to minimize hyperglycaemia and hypoglycaemia –and, hence, to increase time in range– in individuals with diabetes who require insulin treatment is a closed-loop insulin delivery system (also known as the artificial pancreas). Such a system integrates continuous glucose monitoring (CGM), insulin (with or without glucagon) infusion, and a control algorithm to continuously regulate blood glucose levels[6,7]. Multiple studies have shown the merit of incorporating the artificial pancreas into clinical care of individuals with type 1 or type 2 diabetes[8,9].

Despite prior efforts, there are still numerous points that need to be addressed in order to improve the individual components of closed-loop systems[6,10]. With regard to CGM, this includes overcoming sensor delay (i.e., the inherent ~10-minute discrepancy between interstitially measured and actual plasma glucose values), and sensor malfunctions (i.e., periods during which no glucose values are recorded)[6,10,11]. Continuous glucose prediction is a potentially viable strategy to both handle sensor delay and bridge periods of sensor malfunction. The use of machine learning has yielded encouraging glucose prediction accuracy results in relatively small study populations (mostly individuals with type 1 diabetes) or in silico studies, as extensively reviewed elsewhere[12]. Large, human-based study populations are now needed to reliably assess to what extent and within what time interval (i.e., prediction horizon) glucose values can be accurately predicted by use of machine learning. Additionally, incorporation of physical activity, which is considered an important factor for glucose control in daily life, could further improve glucose prediction[6].

In this study, we investigated to what extent glucose values can be accurately predicted at intervals of 15 and 60 minutes by a machine learning model that has been trained with a sliding time window of glucose values preceding the predicted values at a fixed interval. Additionally, we studied whether glucose prediction can be further improved by incorporation of accelerometer-measured physical activity, and to what extent the results differ in a subgroup analysis of individuals with type 2 diabetes only. For this, we used a large population of individuals with either normal glucose metabolism (NGM), prediabetes, or type 2 diabetes who simultaneously underwent CGM and continuous accelerometry during a one-week period. Last, we used the publicly available OhioT1DM Dataset to explore whether CGM-based prediction models would translate to individuals with type 1 diabetes, the primary target population for closed-loop insulin delivery.

Methods

Study population and design

We used data from The Maastricht Study, an observational, prospective, population-based cohort study. The rationale and methodology have been described previously[13]. In brief, The Maastricht Study focuses on the aetiology, pathophysiology, complications and comorbidities of type 2 diabetes, and is characterized by an extensive phenotyping approach. All individuals aged between 40 and 75 years and living in the southern part of the Netherlands were eligible for participation. Participants were recruited through mass media campaigns and from the municipal registries and the regional Diabetes Patient Registry via mailings. For reasons of efficiency, recruitment was stratified according to known type 2 diabetes status, with an oversampling of individuals with type 2 diabetes. In general, the examinations of each participant were performed within a time window of three months. From 19 September 2016 until 13 September 2018, participants were invited to also undergo CGM[14]. During this period, a selected group of recently included participants were invited to return for CGM. In these participants only, there was a median time interval of 2.1 years between CGM and all other measurements. The present report includes cross-sectional data of the 851 participants who had at least 48h of CGM data available and were classified with NGM, prediabetes, or type 2 diabetes. The Maastricht Study has been approved by the institutional medical ethical committee (NL31329.068.10) and the Minister of Health, Welfare and Sports of the Netherlands (Permit 131088-105234-PG). All participants gave written informed consent.

Continuous glucose monitoring

The rationale and methodology of CGM (iPro2 and Enlite Glucose Sensor; Medtronic, Tolochenaz, Switzerland) have been described previously[14]. In brief, the CGM device was worn abdominally and recorded subcutaneous interstitial glucose values (range: 2.2 - 22.2 mmol/L) every five minutes for a seven-day period. For calibration purposes, participants were asked to perform self-measurements of blood glucose four times daily (Contour Next; Ascensia Diabetes Care, Mijdrecht, the Netherlands). Participants were blinded to the CGM recording, but not to self-measured values. Diabetes medication use was allowed and no dietary instructions were given. We only included individuals with at least 48h of CGM, but excluded the first 24h of CGM from analysis because of insufficient calibration. For the glucose prediction analyses, all remaining glucose data points were used. We additionally calculated mean sensor glucose, standard deviation (SD), and coefficient of variation (CV) with the use of Glycemic Variability Research Tool (GlyVaRT; Medtronic) software.

Accelerometry

As described previously, daily physical activity was measured with use of the triaxial activPAL3 accelerometer (PAL technologies; Glasgow, United Kingdom)[13,15]. The accelerometer was, just as the CGM device, attached during the first research visit; participants wore the accelerometer on the front of the right thigh for eight consecutive days. No physical activity instructions were given. PAL Software Suite version 8 (PAL technologies) was used to convert the event-based accelerometry data files into 15-second interval data files. We used the composite of X, Y, and Z accelerations for each 15-second interval as the measure of physical activity.

Assessment of participant characteristics

As described previously[13], we classified glucose metabolism status (GMS) as either NGM, prediabetes, or type 2 diabetes based on both a standardized 2-hour 75 gram oral glucose tolerance test and use of glucose-lowering medication[16]. We assessed medication use as part of a medication interview. Additionally, we determined smoking status and history of diabetes based on questionnaires, measured weight and height –to calculate body mass index (BMI)– and office blood pressure during a physical examination, and measured HbA1c as well as lipid profile in fasting venous blood.

Dataset construction

An overview of data preprocessing, model development, and model evaluation is given in Figure 1. In order to train our models in predicting future glucose values, we constructed two separate datasets (Figure 1, panel a). The first dataset consisted of only the participants' six-day, five-minute interval CGM data (n= 851). The second dataset consisted of both CGM and accelerometry data (n= 540). To synchronize CGM (determined at 5-minute intervals) and accelerometry data (determined at 15-second intervals) in the second dataset, we linearly interpolated glucose values between two glucose data points with a frequency of 15 seconds. Consistent and aligned frequency intervals across these parameters are a statistical precondition for this type of model development[17]. The study populations were randomly split into a training (70%), tuning (10%), and evaluation (20%) dataset such that data from a given individual were present only in one set. The training set was used to train the proposed models. The tuning set was used to iteratively improve the models by selecting the best model architectures and hyperparameters. Finally, the best models were evaluated on the independent evaluation set that was retained during model development.

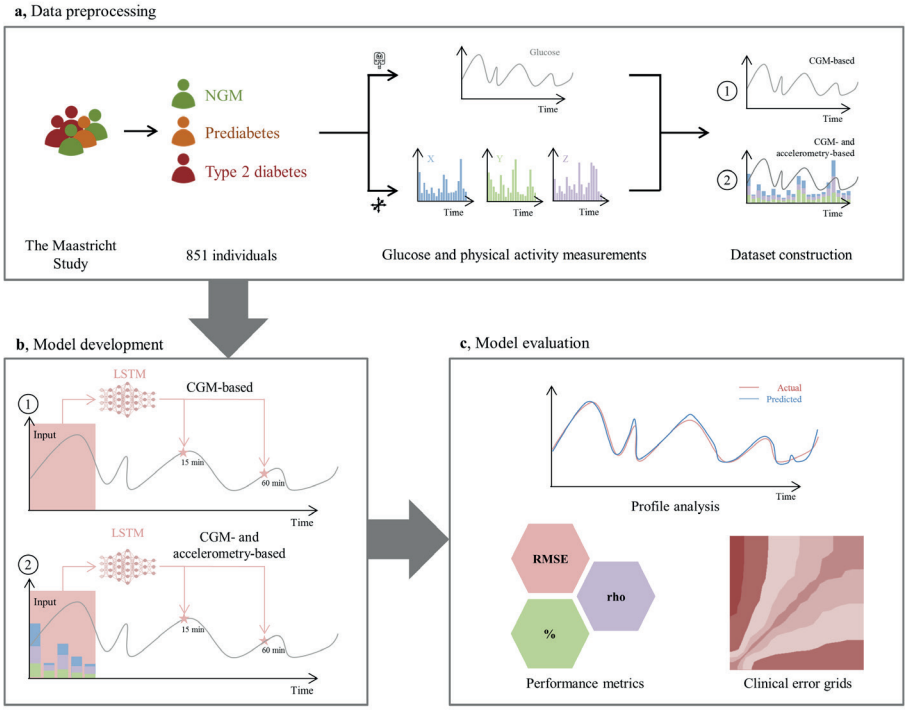


Figure 1. Overview of data preprocessing, model development and evaluation. Data was used from The Maastricht Study, an observational population-based cohort that comprises individuals with normal glucose metabolism (NGM), prediabetes, or type 2 diabetes (panel A). We included 851 individuals who underwent continuous glucose monitoring (CGM), most of whom simultaneously wore an accelerometer to assess physical activity (X, Y, and Z accelerations). Models developed with the long-short term memory (LSTM) architecture were trained in predicting glucose levels at 15- and 60-minute intervals with either CGM data only (1) or both CGM and accelerometer data (2) (panel B). Finally, model performance was evaluated by glucose profile analysis, performance metrics (root-mean-square error [RMSE]; Spearman's correlation coefficient [rho]; proportions), and clinical error grids (panel C).

Model development and design

Our proposed predictive model operates sequentially over CGM and accelerometry data (Figure 1, panel b). At each individual time point, 30 minutes of prior time series data were provided to the statistical model (e.g., six CGM-based glucose values), based on which it predicted glucose values at specified time intervals. For this study, we set these time intervals at 15 and 60 minutes. The nature of this prediction task can be solved by a variety of statistical and machine learning models. In the current study, we assessed autoregressive integrated moving average, support vector regression, gradient-boosting systems, shallow and deep multi-layer perceptron neural networks, and several recurrent neural network (RNN) architectures, including classical RNN[18,19], gated recurrent units[20], long-short term memory (LSTM) networks[21], and all of its bi-directional variants[22,23] (S1 Supporting Information).

Model selection and training

The classical RNN architecture had superior performance at the 15-minute prediction interval (Table 1, RMSE: 0.485 [0.481-0.490]), whilst the LSTM network outperformed all other architectures at the 60-minute prediction interval (Table 1, RMSE: 0.941 [0.937-0.945]). Considering the performance of the LSTM network at a 15-minute prediction interval was nearly as good as the classical RNN, we selected the multi-task LSTM network among several alternatives as architecture of choice to continue our investigations(S1 Supporting Information and Table 1). This architecture runs sequentially over time series data and is able to implicitly model the historical context of an individual by modifying an internal state through time. Specifically, we designed this architecture to predict both time intervals simultaneously, often referred to as “multi-task learning”, which aims to share knowledge amongst prediction tasks.

Next, we evaluated a broad spectrum of hyperparameter combinations for this network (S1 Table). This resulted in a multi-task LSTM architecture, consisting of three layers, including a dropout layer with a total of 56-104 neurons (S2 Table). During training, we used exponential learning-rate decay via the Adam optimization scheme[24]. The best validation results were achieved by use of an initial learning rate with a decay of 0.001 every 1,000 training steps, with a batch size of 1024, and a back-propagation through a time window of 30 minutes. This defines the amount of historic data the model uses, which in our case translates to six (first dataset) or 120 (second dataset) glucose data points, for the model to provide a prediction. The loss function during training was the mean average of the mean-squared error function of all predictions. The maximum amount of epochs was 50.000 with an early stopping criterion (based on 20% hold-out data) set to 250 epochs. We performed data preprocessing, model development, selection, and training using Python programming language (version 3.7.1) with the use of packages Numpy (version 1.17), Pandas (version 0.24), Keras (version 2.2.2), Scikit-learn (version 0.22.0) and Tensorflow (version 2.0.1, beta).

Table 1. Baseline statistical and machine learning model comparison for predicting glucose values.

Prediction window and baseline model Rho		CGM-based glucose prediction		Combined glucose prediction	
		RMSE, mmol/L	Rho	RMSE, mmol/L	
15 minutes	ARIMA	0.842 [0.837 – 0.848]	0.504 [0.490 – 0.518]	0.834 [0.829 – 0.840]	0.498 [0.492 – 0.505]
	SVR	0.791 [0.781 – 0.802]	0.558 [0.549 – 0.567]	0.703 [0.694 – 0.712]	0.612 [0.601 – 0.622]
	LightGBM	0.783 [0.767 – 0.795]	0.589 [0.577 – 0.601]	0.783 [0.771 – 0.794]	0.497 [0.582 – 0.613]
	Shallow MLP	0.810 [0.804 – 0.816]	0.517 [0.506 – 0.529]	0.763 [0.754 – 0.772]	0.592 [0.581 – 0.603]
	Deep MLP	0.807 [0.797 – 0.818]	0.511 [0.504 – 0.518]	0.828 [0.819 – 0.837]	0.510 [0.503 – 0.517]
	RNN	**0.894 [0.887 – 0.902]**	**0.485 [0.481 – 0.490]**	**0.890 [0.882 – 0.898]**	**0.477 [0.472 – 0.482]**
	LSTM	0.872 [0.865 – 0.879]	0.482 [0.477 – 0.487]	0.884 [0.878 – 0.890]	0.501 [0.496 – 0.506]
60 minutes	ARIMA	0.307 [0.284 – 0.329]	1.543 [1.489 – 1.623]	0.303 [0.283 – 0.322]	1.502 [1.455 – 1.568]
	SVR	0.388 [0.376 – 0.398]	1.386 [1.322 – 1.452]	0.394 [0.382 – 0.405]	1.412 [1.350 – 1.475]
	LightGBM	0.500 [0.491 – 0.508]	1.118 [1.098 – 1.136]	0.498 [0.485 – 0.511]	1.128 [1.107 – 1.148]
	Shallow MLP	0.503 [0.495 – 0.511]	1.081 [1.074 – 1.088]	0.483 [0.470 – 0.495]	1.081 [1.070 – 1.092]
	Deep MLP	0.496 [0.484 – 0.509]	1.108 [1.100 – 1.115]	0.515 [0.502 – 0.528]	1.108 [1.099 – 1.017]
	RNN	0.591 [0.581 – 0.600]	0.989 [0.983 – 0.995]	0.596 [0.589 – 0.603]	0.992 [0.984 – 0.998]
	LSTM	**0.605 [0.593 – 0.616]**	**0.941 [0.937 – 0.945]**	**0.602 [0.595 – 0.609]**	**0.922 [0.919 – 0.926]**

Performance was assessed by Spearman's rank correlation coefficient (rho) and root-mean-square error (RMSE). Data are reported as median [95% confidence intervals], calculated using 1,000 bootstraps.

Translation of the prediction models to the OhioT1DM Dataset

We used data from the OhioT1DM Dataset to explore whether our CGM-based prediction models would translate to individuals with type 1 diabetes. The OhioT1DM Dataset is freely available for scientific purposes and contains data of 6 individuals with type 1 diabetes who were all using insulin pump therapy and CGM[25]. The participants provided interstitial glucose values every five minutes for an eight-week period. First, in order to also include 30-minute prediction, we retrained our main CGM-based models on the main study population with identical hyperparameters and settings (S2 Table). Then, we evaluated the main CGM-based model on the test portion of the OhioT1DM Dataset (20%). Next, we aimed to optimize our main CGM-based model by training it on the train portion of the OhioT1DM Dataset. Specifically, we trained the model using an Adam optimizer with a learning rate of 10-4, a batch size of 1024, a maximum of 10.000 epochs and an early stopping criterion (based on 20% of the training data) set to 100 epochs. Last, we evaluated this optimized model on the test portion using performance metrics and safety error grids, as described previously.

Model evaluation and statistical analysis

Model evaluation was performed in the independent evaluation sets of individuals that were not used during model development (Figure 1, panel c). We employed several metrics to assess the performance of our models: root-mean-square error (RMSE), proportion of predicted values within 5% or 10% of actual glucose values, and Spearman's rank correlation coefficient (rho) (S2 Supporting Information). Bootstrapping was performed to obtain 95% confidence intervals for each of these metrics[26]. In addition, we used error grids that are classically used for assessment of blood glucose monitor safety (i.e., surveillance error grid, Parkes error grid) to evaluate the safety of our glucose prediction models[27,28]. Last, we performed several sensitivity analysis in our main study population by stratifying model performance for: (1) GMS (i.e., separate results for NGM and prediabetes); (2) day (06.00 to 24.00h) and night (24.00 to 06.00h); and (3) low or high glucose variability, defined as the 97.5th percentile of CGM-assessed SD in individuals with NGM (SD > 1.37 mmol/L)[14].

Normally distributed data are presented as mean ± SD, non-normally distributed data as median and interquartile range, and categorical data as n (%). Statistical analyses were performed using the Statistical Package for Social Sciences (version 25.0; IBM, Chicago, Illinois, USA) and the Python programming language (version 3.7.1).

Results

Main study population characteristics

In total, 896 individuals underwent CGM as part of The Maastricht Study's extensive phenotyping approach. We included participants with at least 48h of CGM data and either NGM, prediabetes, or type 2 diabetes. This resulted in the final study population of 851 individuals. Of this population, 540 participants (63.5%) simultaneously underwent CGM and accelerometry.

Table 2 shows the overall and type 2 diabetes-stratified characteristics of the two study populations (CGM-based as well as CGM- and accelerometry-based glucose prediction). The overall participant characteristics of both populations were generally comparable with regard to age, sex, BMI, glycaemic indices, blood pressure, and lipid profile, although the latter contained fewer participants with prediabetes or type 2 diabetes. Additionally, the participants with type 2 diabetes in the CGM- and accelerometry-based glucose prediction population were more often newly diagnosed with type 2 diabetes. Accordingly, these participants less often used glucose-lowering medication. Participant characteristics of the NGM and prediabetes subgroups are described in S3 Table.

Table 2. Participant characteristics of the CGM-based and CGM- and accelerometry-based glucose prediction study populations.

	CGM-based glucose prediction		**CGM- and accelerometry-based glucose prediction**	
Characteristic	**Total (n=851)**	**T2D (n=197)**	**Total (n=540)**	**T2D (n=68)**
Age, years	59.9 ± 8.7	62.4 ± 7.8	59.1 ± 8.7	62.0 ± 6.9
Women, n (%)	418 (49.1)	69 (35.0)	276 (51.1)	22 (32.4)
BMI, kg/m^2	27.2 ± 4.4	29.7 ± 4.7	26.5 ± 4.0	28.6 ± 4.1
Newly diagnosed T2D, n (%)	70 (8.2)	70 (35.5)	35 (6.5)	35 (51.5)
Glucose metabolism status				
NGM/PreD/T2D, n	470/184/197	-	372/99/68	-
NGM/PreD/T2D, %	55.2/21.6/23.1	-	69.1/18.3/12.6	-
Fasting plasma glucose, mmol/L	5.4 [5.0 – 6.2]	7.3 [6.5 – 8.4]	5.3 [4.9 – 5.8]	7.2 [6.3 – 8.4]
2-h post-load glucose, mmol/L	6.7 [5.2 – 9.1]	13.6 [11.7 – 16.2]	6.2 [5.0 – 7.7]	12.5 [11.3 – 16.6]
HbA_{1c}, %	5.7 ± 0.8	6.7 ± 1.0	5.6 ± 0.6	6.4 ± 0.9
HbA_{1c}, mmol/mol	39.1 ± 8.3	49.2 ± 10.8	37.3 ± 6.2	46.9 ± 10.2
Sensor glucose Mean, mmol/L	6.1 [5.7 – 6.7]	7.5 [6.8 – 8.7]	5.9 [5.6 – 6.4]	7.3 [6.5 – 8.2]
SD, mmol/L	0.84 [0.68 – 1.18]	1.51 [1.14 – 1.95]	0.79 [0.66 – 1.01]	1.46 [0.94 – 1.99]
SD > 1.37 mmol/L, n (%)	142 (16.7)	115 (58.4)	50 (9.3)	36 (52.9)
CV, %	14.0 [11.6 – 17.6]	19.3 [15.9 – 24.0]	13.3 [11.2 – 16.8]	19.2 [14.5 – 24.1]
Diabetes medication use, n (%)	109 (12.8)	109 (55.6)	27 (4.8)	27 (39.7)
Insulin	19 (2.2)	19 (9.6)	4 (0.7)	4 (5.9)
Metformin	104 (12.2)	104 (53.1)	27 (5.0)	27 (39.7)
Sulfonylureas	21 (2.5)	21 (10.7)	6 (1.1)	6 (8.8)
Thiazolidinediones	0 (0)	0 (0)	0 (0)	0 (0)
GLP-1 analogues	3 (0.4)	3 (1.5)	1 (0.2)	1 (1.5)
DDP-4 inhibitors	1 (0.1)	1 (0.5)	0 (0)	0 (0)
SGLT-2 inhibitors	1 (0.1)	1 (0.5)	0 (0)	0 (0)
Office SBP, mmHg	133.3 ± 18.0	139.4 ± 15.6	132.2 ± 17.9	137.7 ± 15.3
Office DBP, mmHg	75.2 ± 10.2	77.7 ± 10.5	74.7 ± 10.1	77.7 ± 9.6
Antihypertensive medication use, n (%)	305 (35.9)	126 (64.3)	162 (30.0)	41 (60.3)
Total-to-HDL cholesterol ratio	3.5 [2.8 – 4.3]	3.6 [2.9 – 4.3]	3.4 [2.8 – 4.3]	3.7 [2.8 – 4.6]
Triglycerides, mmol/L	1.3 [0.9 – 1.8]	1.5 [1.0 – 2.1]	1.2 [0.9 – 1.7]	1.6 [1.0 – 2.3]
Lipid-modifying medication use, n (%)	212 (24.9)	115 (58.4)	100 (18.5)	39 (57.4)
Smoking status				
Never/former/current, n	327/415/106	67/104/26	214/253/70	19/36/13
Never/former/current, %	38.6/48.9/12.5	34.0/52.8/13.2	39.9/47.1/13.0	27.9/52.9/19.1

Data are reported as mean ± SD, median [interquartile range], or number (percentage [%]) as appropriate. CGM, continuous glucose monitoring; BMI, body mass index; T2D, type 2 diabetes; NGM, normal glucose metabolism; PreD, prediabetes; HbA1c, glycated haemoglobin A1c; SD, standard deviation; CV, coefficient of variation; GLP-1, glucagon-like peptide-1; DPP-4, dipeptidase-4; SGLT-2, sodium-glucose cotransporter 2; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL, high-density lipoprotein.

Overall performance of machine learning-based glucose prediction

We trained two machine learning models (i.e., CGM-based; CGM- and accelerometry-based) in predicting glucose levels at 15- and 60-minute intervals. Visually, both models appeared capable of accurately predicting the real glucose profiles, as illustrated by the representative examples in S1 Figure and S2 Figure. Next, we assessed the performance of our models in our evaluation datasets with a variety of metrics, including an average error term (RMSE), the proportion of predictions within 5% or 10% deviation of the actual value, and correlation (rho). The evaluation datasets comprise 20% of the original or stratified study populations and thus vary in sample size (n= 13 - 170).

Overall, our models demonstrated high prediction accuracy, supported by low RMSE values and high proportions of predicted glucose values within 5% and 10% deviation (Table 3). Model performance in the type 2 diabetes subgroup was generally lower compared to the overall group, except for correlation coefficients, which were often higher in individuals with type 2 diabetes. This phenomenon can be largely attributed to the lower correlation coefficients of individuals with NGM and prediabetes (S4 Table), which is caused by range restriction (i.e., smaller glucose ranges attenuate the correlation coefficients)[29]. Consequently, the correlation coefficients are valid for the comparison of CGM-based glucose prediction to CGM- and accelerometry-based glucose prediction, but not for comparison of the overall study population to the type 2 diabetes subgroup. In addition, we observed short-to-moderate time lags for the 15- and 60-minute predictions (S5 Table).

In general, incorporation of accelerometry data in the models only slightly improved performance metrics at both prediction intervals (Table 3). S4 Table shows the model performance in NGM and prediabetes subgroups. Glucose prediction was most precise in individuals with NGM. Of note, the ML-based models substantially outperformed a naive approach that used t0 as predicted glucose value (S6 Table, S3 and S4 Figures).

Table 3. Overall performance in the main study population of CGM-based and CGM- and accelerometry-based machine learning models trained in predicting glucose values at time intervals of 15 and 60 minutes.

		CGM-based glucose prediction		CGM- and accelerometry-based glucose prediction	
		Total (n=170)	Total (n=43)	T2D (n=109)	T2D (n=13)
15 minutes	RMSE, mmol/L	0.188 [0.186 – 0.191]	0.288 [0.281 – 0.306]	0.184 [0.177 – 0.189]	0.271 [0.260 – 0.282]
	< 5% , %	92.98 [92.87 – 93.05]	92.02 [91.83 – 92.25]	93.06 [93.03 – 93.09]	92.04 [91.99 – 92.11]
	< 10% , %	99.17 [99.13 – 99.23]	98.88 [98.82 – 98.94]	99.25 [99.21 – 99.28]	98.90 [98.83 – 98.97]
	Rho	0.961 [0.959 – 0.962]	0.987 [0.985 – 0.989]	0.968 [0.964 – 0.970]	0.990 [0.988 – 0.993]
60 minutes	RMSE, mmol/L	0.589 [0.582 – 0.592]	0.701 [0.692 – 0.711]	0.582 [0.579 – 0.586]	0.700 [0.693 – 0.708]
	< 5% , %	70.22 [70.09 – 70.41]	66.23 [66.13 – 66.33]	70.11 [70.05 – 70.17]	66.17 [66.09 – 66.22]
	< 10% , %	87.39 [87.24 – 87.53]	85.82 [85.70 – 85.93]	87.44 [87.38 – 87.50]	86.11 [86.01 – 86.20]
	Rho	0.721 [0.719 – 0.722]	0.781 [0.779 – 0.782]	0.725 [0.721 – 0.729]	0.790 [0.782 – 0.799]

Data are reported as mean [95% confidence interval]. CGM, continuous glucose monitoring; T2D, type 2 diabetes; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

Safety evaluation with clinical error grids

We assessed the safety of our machine learning-based glucose prediction using two clinical error grids (i.e., surveillance and Parkes error grids). Figure 2 depicts the safety results for individuals with type 2 diabetes according to the surveillance error grid. At the 15-minute interval, almost all predictions (>99.9%) were clinically safe (i.e., a risk score between 0 and 1.0) (Figure 2, panels A and B). At the extended prediction window of 60 minutes, clinical safety was slightly lower (98.4-99.2%) (Figure 2, panels C and D). Parkes error grid assessment yielded similar results (S5 Figure). Of note, less accurate predictions were more often in the vertical B-D zones than in the horizontal B-E zones (e.g., S4 Figure, panel C: 11.80% versus 4.24%), which suggests a model tendency to underestimate rather than overestimate actual glucose values, the latter of which being more dangerous.

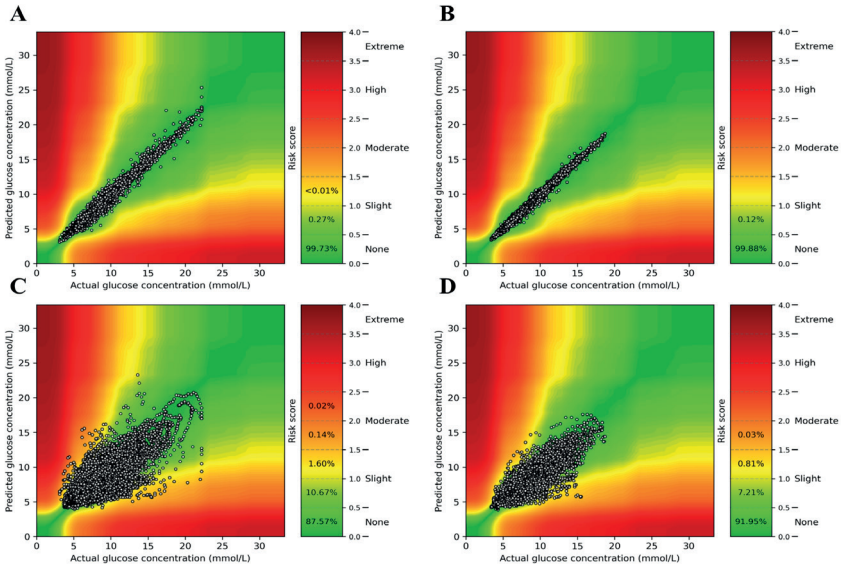


Figure 2. Surveillance error grid evaluation of glucose prediction safety at time intervals of 15 and 60 minutes in the main study population. Assessment of CGM-based glucose prediction safety in individuals with type 2 diabetes (n=43) at 15 minutes (panel A) and 60 minutes (panel C). Assessment of CGM- and accelerometry-based glucose prediction safety in individuals with type 2 diabetes (n=13) at 15 minutes (panel B) and 60 minutes (panel D). The risk score values translate to the following degrees of risk: 0 - 0.5, none; 0.5 - 1.0, slight (lower); 1.0 - 1.5, slight (higher); 1.5 - 2.0, moderate (lower); 2.0 - 2.5, moderate (higher); 2.5 – 3.0, great (lower); 3.0 - 3.5, great (higher); > 3.5 extreme[27].

Additional analyses

To further obtain insights into our model predictions, we assessed performance metrics stratified by day and night (S7 Table). Fifteen-minute predictions did not materially differ between day and night. By contrast, accuracy of 60-minute predictions was lower during the day than at night. In addition, we stratified the results by high or low glucose variability (i.e., SD cut-off of 1.37 mmol/L) (S8 Table). Model performance was slightly lower at higher glucose variability, at both time intervals of 15 and 60 minutes.

Translation of the prediction models to the OhioT1DM Dataset

The prediction accuracy of the CGM-based model that was developed with our main study population was moderate in individuals with type 1 diabetes (RMSEs at 15, 30, and 60 min: 0.689 [0.685 – 0.693], 1.189 [1.183 – 1.195], and 1.918 [1.910 – 1.926] mmol/L), but substantially improved after being trained on data from each individual with type 1 diabetes (RMSEs at 15, 30, and 60 min: 0.426 [0.422 – 0.430], 1.046 [1.039 – 1.052], and 1.733 [1.725 – 1.741] mmol/L; S9 Table). Accordingly, clinical safety was substantial as shown by the high percentages of clinically safe predictions (15-minute: >99%, 30-minute: >97%, and 60-minute: >91%; Figure 3).

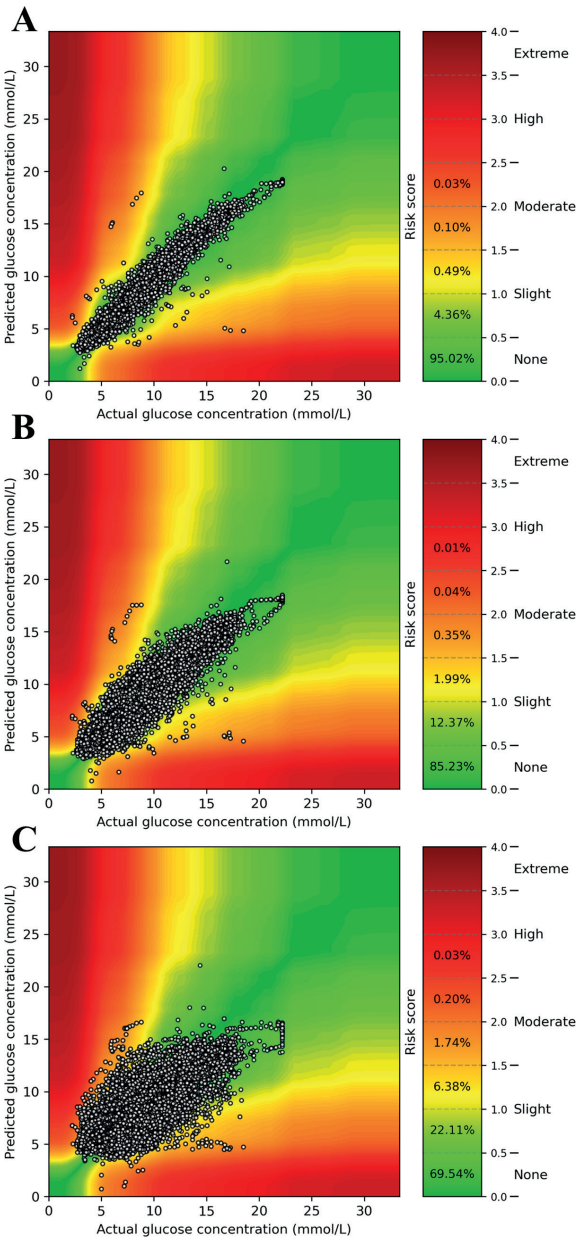


Figure 3. Surveillance error grid evaluation of glucose prediction safety at time intervals of 15, 30, and 60 minutes in individuals with type 1 diabetes. Assessment of CGM-based glucose prediction safety in individuals with type 1 diabetes (n=6) at 15 (panel A), 30 (panel B), and 60 minutes (panel C). The risk score values translate to the following degrees of risk: 0 - 0.5, none; 0.5 - 1.0, slight (lower); 1.0 - 1.5, slight (higher); 1.5 - 2.0, moderate (lower); 2.0 - 2.5, moderate (higher); 2.5 – 3.0, great (lower); 3.0 - 3.5, great (higher); > 3.5 extreme[27].

Discussion

In this study with 851 individuals and almost 1.4 million glucose measurements, we investigated whether glucose values can be accurately predicted by using machine learning-based models that utilise recently measured CGM and physical activity data with the prospect of improving closed-loop insulin delivery systems. Our study has several important findings and unique characteristics. First, the machine learning-based models are capable of accurately predicting the actual glucose profiles at 15 minutes, as reflected by several objective performance metrics (e.g., RMSE, rho; Table 2) and visual illustrations (S1 Figure and S2 Figure). Despite prediction accuracy being moderately lower at 60 minutes, more than 98% of the predicted values remained sufficiently accurate to be deemed clinically safe based on surveillance error grids (Figure 2). Second, glucose prediction only improved slightly when accelerometer-assessed physical activity data was incorporated in the models. Third, translation of our CGM-based glucose prediction models to individuals with type 1 diabetes yielded encouraging results (i.e., ample prediction accuracy and clinical safety).

Although most research has thus far focused on type 1 diabetes[12], several efforts have been made to use machine learning for glucose prediction in individuals with type 2 diabetes[30-34]. Most of these studies assessed technical aspects of glucose prediction in relatively small (n=1 to 50) or even virtual, in silico populations. Such studies provide valuable comparisons of models, but show suboptimal and highly variable performance in predicting glucose values. To our knowledge, this is the first study to report this level of performance in a large, population-based sample of individuals with NGM, prediabetes, or type 2 diabetes. Our CGM-based models were able to accurately predict glucose values at 15 (RMSEs, overall/type 2 diabetes: 0.19/0.29 mmol/L) and 60 minutes (RMSEs, overall/type 2 diabetes: 0.59/0.70 mmol/L). These results surpass previously reported RMSE values for a sample of 50 individuals with type 2 diabetes, which were 0.65 and 1.50 mmol/L for 15- and 60-minute CGM-based glucose prediction, respectively[34]. We expect this difference to, in part, stem from our much larger sample size. To our knowledge, our exploratory translation to individuals with type 1 diabetes (S9 Table) showed that our models perform equally well as recent publications in the field[12,35-38]. For example, the best performing model of the Blood Glucose Level Prediction Challenge 2018, which was also based on a LSTM architecture as well as was trained on and evaluated in the OhioT1DM Dataset, reported 30-minute and 60-minute RMSEs of 1.05 and 1.74 mmol/L[35]. Additionally, Kriventsov et al. recently described large-scale application of glucose prediction in a smartphone app (Diabits) and reported a comparable RMSE at 30 minutes (1.04 mmol/L)[36]. We anticipate that further technical development of our prediction models, while using a larger sample of individuals with type 1 diabetes, will advance performance even more.

We integrated physical activity, which we assessed via accelerometry, into our glucose prediction model, because of its short- and long-term effects on daily glucose patterns. Whereas an acute bout of physical activity can either decrease or increase serum glucose levels, prolonged exercise improves insulin sensitivity, and thus insulin-stimulated glucose uptake[39]. While it should be noted that CGM- and accelerometry-based glucose prediction yielded larger improvements relative to CGM-based glucose prediction for the 60-minute interval, most notably during the day (S7 Table) and in individuals with higher glucose variability (S9 Table), incorporation of physical activity generally only marginally improved glucose prediction. This can be explained by the observation that the models based on CGM data only already performed very well, which limits the ability to achieve additional improvements[40]. Also, the effect of physical activity on serum glucose levels is relatively small in people with beta-cell function that is either normal or only mildly deficient. Given the absence of pancreatic glucoregulation in individuals with type 1 diabetes, it is conceivable that incorporation of accelerometry data leads to more substantially improved model performance in this patient group[40], which, at present, we were not able to further explore. In addition, a time interval of 15 or 60 minutes could be too short to incorporate long-term physical activity effects into the prediction model.

The closed-loop insulin delivery system has been shown to improve glycaemic control in individuals with type 1 or type 2 diabetes[8,9,41]. Nevertheless, several aspects of the artificial pancreas require further enhancement[6,10]. Our results demonstrate that machine learning-based glucose prediction has the promise of being a valid and safe strategy to both overcome ~10-minute sensor delay and bridge prolonged periods of sensor malfunction. Not only are more than 99% of the predicted glucose values in clinically safe zones (i.e., Parkes error grid zone A and B), the model also tended to slightly underestimate rather than overestimate the actual glucose values. In case the prediction model were to be implemented, this would further reduce the risk of iatrogenic hypoglycaemia. Nevertheless, future research is needed to assess whether incorporation of these prediction models in a closed-loop insulin delivery system safely improves glycaemic control.

This proof-of-principle study has several strengths and limitations. Strengths are 1) the largest well-characterized, population-based study sample thus far, which ensured sufficient statistical power; 2) the unique large-scale combination of CGM and continuous accelerometry, which enabled us to study to what extent incorporation of data on physical activity would improve prediction in this population; 3) the gold-standard assessment of GMS, which allowed for the comparison of performance in NGM, prediabetes and type 2 diabetes; 4) the broad and solid evaluation of various statistical and machine learning architectures for this prediction task; and 5) result robustness, as reflected by the

consistency of several statistical and clinical performance metrics.

Our research had certain limitations. First, the main study population comprised individuals with NGM, prediabetes, or type 2 diabetes, who are generally not the target population for closed-loop insulin delivery systems. We, therefore, exploratively investigated whether our prediction models would translate to individuals with type 1 diabetes using the OhioT1DM Dataset, which yielded encouraging results. Nevertheless, we underscore the importance of extensive evaluation of the models in a larger sample of individuals with type 1 diabetes, insulin-treated type 2 diabetes, or both. Second, we were unable to factor in other important elements pertaining to glycaemic control (e.g., diet or medication use)[6]. In automated, self-regulatory closed-loop systems, utilization of these kinds of data requires manual input, which is less convenient and reliable than CGM. In addition, since glucose prediction was only slightly improved by incorporating physical activity, we expect relatively little gain from including such factors into our models, at least in individuals with type 2 diabetes. However, given the results of several small studies that have incorporated diet and medication use[12], we acknowledge that this may not hold true for individuals with type 1 diabetes. In this regard, large-scale studies are required to reach more definitive conclusions. If diet, medication use, or other factors were to be incorporated, it is necessary to evaluate whether LSTM remains the best-performing machine learning architecture.

Conclusion

In this study, we show that our machine learning-based models are able to accurately and safely predict glucose values for up to 60 minutes in individuals with, NGM, prediabetes, or type 2 diabetes. In addition, translation of our prediction models to individuals with type 1 diabetes showed encouraging results. We observed particularly high precision at a 15-minute prediction window, which is a clinically relevant timespan to align interstitially measured glucose values by continuous glucose measurement systems with actual plasma glucose values. As such, the prediction model can be used to improve closed-loop insulin delivery systems by overcoming sensor delay. In addition, longer prediction intervals may be used to safely bridge periods of sensor malfunction. Last, our current findings question the use of accelerometry to substantially improve prediction. Future research should validate our findings by replicating the results in a larger sample of individuals with type 1 diabetes and studying the effects of implementing the prediction model in a closed-loop insulin delivery system.

References

1. Collaboration NCDRF. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. Lancet. 2016;387(10027):1513-30. doi: 10.1016/S0140-6736(16)00618-8. PubMed PMID: 27061677; PubMed Central PMCID: PMCPMC5081106.
2. Emerging Risk Factors C, Sarwar N, Gao P, Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. Lancet. 2010;375(9733):2215-22. doi: 10.1016/S0140-6736(10)60484-9. PubMed PMID: 20609967; PubMed Central PMCID: PMCPMC2904878.
3. Forbes JM, Cooper ME. Mechanisms of diabetic complications. Physiol Rev. 2013;93(1):137-88. doi: 10.1152/physrev.00045.2011. PubMed PMID: 23303908.
4. American Diabetes A. 6. Glycemic Targets: Standards of Medical Care in Diabetes-2019. Diabetes Care. 2019;42(Suppl 1):S61-S70. Epub 2018/12/19. doi: 10.2337/dc19-S006. PubMed PMID: 30559232.
5. International Hypoglycaemia Study G. Hypoglycaemia, cardiovascular disease, and mortality in diabetes: epidemiology, pathogenesis, and management. Lancet Diabetes Endocrinol. 2019;7(5):385-96. Epub 2019/03/31. doi: 10.1016/S2213-8587(18)30315-2. PubMed PMID: 30926258.
6. Cobelli C, Renard E, Kovatchev B. Artificial pancreas: past, present, future. Diabetes. 2011;60(11):2672-82. Epub 2011/10/26. doi: 10.2337/db11-0654. PubMed PMID: 22025773; PubMed Central PMCID: PMCPMC3198099.
7. Bruttomesso D. Toward Automated Insulin Delivery. N Engl J Med. 2019;381(18):1774-5. Epub 2019/10/17. doi: 10.1056/NEJMe1912822. PubMed PMID: 31618534.
8. Weisman A, Bai JW, Cardinez M, Kramer CK, Perkins BA. Effect of artificial pancreas systems on glycaemic control in patients with type 1 diabetes: a systematic review and meta-analysis of outpatient randomised controlled trials. Lancet Diabetes Endocrinol. 2017;5(7):501-12. Epub 2017/05/24. doi: 10.1016/S2213-8587(17)30167-5. PubMed PMID: 28533136.
9. Kumareswaran K, Thabit H, Leelarathna L, Caldwell K, Elleri D, Allen JM, et al. Feasibility of closed-loop insulin delivery in type 2 diabetes: a randomized controlled study. Diabetes Care. 2014;37(5):1198-203. Epub 2013/09/13. doi: 10.2337/dc13-1030. PubMed PMID: 24026542.
10. Blauw H, Keith-Hynes P, Koops R, DeVries JH. A Review of Safety and Design Requirements of the Artificial Pancreas. Ann Biomed Eng. 2016;44(11):3158-72. Epub 2016/11/04. doi: 10.1007/s10439-016-1679-2. PubMed PMID: 27352278; PubMed Central PMCID: PMCPMC5093196.
11. Rodbard D. Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities. Diabetes Technol Ther. 2016;18 Suppl 2:S3-S13. Epub 2016/01/20. doi: 10.1089/dia.2015.0417. PubMed PMID: 26784127; PubMed Central PMCID: PMCPMC4717493.
12. Woldaregay AZ, Arsand E, Walderhaug S, Albers D, Mamykina L, Botsis T, et al. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. Artif Intell Med. 2019;98:109-34. Epub 2019/08/07. doi: 10.1016/j.artmed.2019.07.007. PubMed PMID: 31383477.
13. Schram MT, Sep SJ, van der Kallen CJ, Dagnelie PC, Koster A, Schaper N, et al. The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities. Eur J Epidemiol. 2014;29(6):439-51. doi: 10.1007/s10654-014-9889-0. PubMed PMID: 24756374.
14. Foreman YD, Brouwers M, van der Kallen CJH, Pagen DME, van Greevenbroek MMJ, Henry RMA, et al. Glucose variability assessed with continuous glucose monitoring: reliability, reference values and correlations with established glycaemic indices - The Maastricht Study. Diabetes Technol Ther. 2019. Epub 2019/12/31. doi: 10.1089/dia.2019.0385. PubMed PMID: 31886732.
15. van der Berg JD, Stehouwer CD, Bosma H, van der Velde JH, Willems PJ, Savelberg HH, et al. Associations of total amount and patterns of sedentary behaviour with type 2 diabetes and the metabolic syndrome: The Maastricht Study. Diabetologia. 2016;59(4):709-18. Epub 2016/02/03. doi: 10.1007/s00125-015-3861-8. PubMed PMID: 26831300; PubMed Central PMCID: PMCPMC4779127.
16. WHO. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF

consultation. WHO. 2006.
17. Staudemeyer RC, Rothstein Morris E. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv e-prints [Internet]. 2019 September 01, 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190909586S.
18. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv e-prints. 2018:arXiv:1808.03314.
19. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533-6. doi: 10.1038/323533a0.
20. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv e-prints. 2014:arXiv:1412.3555.
21. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735-80. doi: 10.1162/neco.1997.9.8.1735.
22. Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM Networks for Improved Phoneme Classification and Recognition2005. 799-804 p.
23. Schuster M, Paliwal K. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on. 1997;45:2673-81. doi: 10.1109/78.650093.
24. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv e-prints [Internet]. 2014 December 01, 2014:[arXiv:1412.6980 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2014arXiv1412.6980K.
25. Marling C, Bunescu RC, editors. The OhioT1DM Dataset For Blood Glucose Level Prediction. KHD@IJCAI; 2018.
26. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York, N.Y.; London: Chapman & Hall; 1993.
27. Klonoff DC, Lias C, Vigersky R, Clarke W, Parkes JL, Sacks DB, et al. The surveillance error grid. J Diabetes Sci Technol. 2014;8(4):658-72. Epub 2015/01/07. doi: 10.1177/1932296814539589. PubMed PMID: 25562886; PubMed Central PMCID: PMCPMC4764212.
28. Pfutzner A, Klonoff DC, Pardo S, Parkes JL. Technical aspects of the Parkes error grid. J Diabetes Sci Technol. 2013;7(5):1275-81. Epub 2013/10/16. doi: 10.1177/193229681300700517. PubMed PMID: 24124954; PubMed Central PMCID: PMCPMC3876371.
29. Bland JM, Altman DG. Correlation in restricted ranges of data. BMJ. 2011;342:d556. doi: 10.1136/bmj.d556. PubMed PMID: 21398359.
30. Sudharsan B, Peeples M, Shomali M. Hypoglycemia prediction using machine learning models for patients with type 2 diabetes. J Diabetes Sci Technol. 2015;9(1):86-90. Epub 2014/10/16. doi: 10.1177/1932296814554260. PubMed PMID: 25316712; PubMed Central PMCID: PMCPMC4495530.
31. Georga E, Protopappas V, Fotiadis D. Glucose Prediction in Type 1 and Type 2 Diabetic Patients Using Data Driven Techniques. 2011.
32. Faruqui SHA, Du Y, Meka R, Alaeddini A, Li C, Shirinkam S, et al. Development of a Deep Learning Model for Dynamic Forecasting of Blood Glucose Level for Type 2 Diabetes Mellitus: Secondary Analysis of a Randomized Controlled Trial. JMIR Mhealth Uhealth. 2019;7(11):e14452. Epub 2019/11/05. doi: 10.2196/14452. PubMed PMID: 31682586; PubMed Central PMCID: PMCPMC6858613.
33. Albers DJ, Levine M, Gluckman B, Ginsberg H, Hripcsak G, Mamykina L. Personalized glucose forecasting for type 2 diabetes using data assimilation. PLoS Comput Biol. 2017;13(4):e1005232. Epub 2017/04/28. doi: 10.1371/journal.pcbi.1005232. PubMed PMID: 28448498; PubMed Central PMCID: PMCPMC5409456.
34. Mohebbi A, Johansen AR, Hansen N, Christensen PE, Tarp JM, Jensen ML, et al. Short Term Blood Glucose Prediction based on Continuous Glucose Monitoring Data. arXiv e-prints [Internet]. 2020 February 01, 2020:[arXiv:2002.02805 p.]. Available from: https://ui.adsabs.harvard.edu/abs/2020arXiv200202805M.
35. Martinsson J, Schliep A, Eliasson B, Mogren O. Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks. Journal of Healthcare Informatics Research. 2020;4(1):1-18. doi: 10.1007/s41666-019-00059-y.
36. Kriventsov S, Lindsey A, Hayeri A. The Diabits App for Smartphone-Assisted Predictive Monitoring of Glycemia in Patients With Diabetes: Retrospective Observational Study. JMIR Diabetes. 2020;5(3):e18660. Epub 2020/09/23. doi: 10.2196/18660. PubMed PMID: 32960180; PubMed Central PMCID: PMCPMC7539161.

37. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: A Deep Learning Framework for Accurate Glucose Forecasting. IEEE J Biomed Health Inform. 2020;24(2):414-23. Epub 2019/08/02. doi: 10.1109/JBHI.2019.2931842. PubMed PMID: 31369390.
38. Chen J, Li K, Herrero P, Zhu T, Georgiou P, editors. Dilated Recurrent Neural Network for Short-time Prediction of Glucose Concentration. KHD@IJCAI; 2018.
39. Stanford KI, Goodyear LJ. Exercise and type 2 diabetes: molecular mechanisms regulating glucose uptake in skeletal muscle. Adv Physiol Educ. 2014;38(4):308-14. Epub 2014/12/01. doi: 10.1152/advan.00080.2014. PubMed PMID: 25434013; PubMed Central PMCID: PMCPMC4315445.
40. Pencina MJ, D'Agostino RB, Pencina KM, Janssens AC, Greenland P. Interpreting incremental value of markers added to risk prediction models. Am J Epidemiol. 2012;176(6):473-81. Epub 2012/08/10. doi: 10.1093/aje/kws207. PubMed PMID: 22875755; PubMed Central PMCID: PMCPMC3530349.
41. Blauw H, Onvlee AJ, Klaassen M, van Bon AC, DeVries JH. Fully Closed Loop Glucose Control With a Bihormonal Artificial Pancreas in Adults With Type 1 Diabetes: An Outpatient, Randomized, Crossover Trial. Diabetes Care. 2021. Epub 2021/01/06. doi: 10.2337/dc20-2106. PubMed PMID: 33397767.

Supplemental material

Supporting information

S1 supporting information. Background information on machine learning models reviewed in current study

We conducted a comparison of available algorithms on the continuous glucose prediction task. We considered the following algorithms:

- ARIMA: an autoregressive integrated moving average (ARIMA) model is an adaptation of the autoregressive moving average (ARMA) model. These models aim to estimate the time series using two polynomials, one for autoregression (AR), and the other for the moving average (MA). ARIMA models for glucose prediction have widely been described in the literature[1-4]. For the current comparison, we used a ARIMA (p=3, d=0, q=0) model based on previous findings by Otoom et al[1].
- Support vector regression: support vector regression (SVR) is a generalization of support vector machines (SVM) that work for regression problems. SVR models for the prediction of glucose values have been described in the literature[5,6]. For the current comparison, we build the SVR model with a Gaussian RFB kernel that was optimized by the differential evolution algorithm as described by Georga et al[5].
- Gradient-boosting trees: gradient-boosting trees have shown superior performances in the medical domain, but lack the capability to aggregate information over time. For baseline comparison, we used the LightGBM[7] implementation with a learning rate of 0.01, a maximum number of trees of 500, and a maximum depth of each base learner to be 50.
- Feed-forward neural networks: feed-forward neural networks do not have the capacity to aggregate the information about past glucose status of an individual over time. Yet, they are relatively simple networks which have previously been employed for glucose[8-10]. We considered shallow and deep multi-layer perceptron (MLP) neural networks. The shallow MLP consisted of one hidden layer (ReLU) with 16 neurons. The deep MLP consisted of three hidden layers (ReLU) with 64, 32, and 16 neurons, respectively. Both networks were trained using the Adam optimizer scheme with a learning rate of 0.001.
- Recurrent neural networks: recurrent neural networks are a type of neural networks with the capacity to explicitly model information over time. These type of neural networks are designed to work with temporal data such as glucose and physical activity data. Although we acknowledge the broad availability of neural network architectures[11-15], we only considered recurrent neural networks (RNN), and long-short term memory (LSTM) networks for the baseline comparison. Both models consisted of one RNN layer (either RNN or LSTM) with 32 neurons, followed by a Dense layer of 8 neurons. Both networks were trained using the Adam optimizer scheme with a learning rate of 0.001.

S2 supporting information. Background information on metrics used in the current study.
In the current study we used several metrics to assess the performance of our models. In this paragraph we will describe each of them briefly and also provide their mathematical definition.

- Root-mean-square error (RMSE): the RMSE is an average error term which is in the order of the predicted / measured variable. It also can be interpreted as the standard deviation of the prediction errors. The formula of RMSE is as follows:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i - y_i\right)^2}{n}}$$

with $\hat{y}_1, \hat{y}_2, \hat{y}_n$ *depicting the predicted values;* y_1, y_2, y_n *depicting the real values and* n *representing the total number of predictions*

Correlation: correlation is a measure of how well the relationship between two variables is. In this study, we deal with non-parametrically distributed data and therefore use Spearman's rank correlation coefficient (rho) which tries to describe the relationship of two variables using a monotonic function:

$$\rho = \frac{S_{xy}}{S_x S_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\left(R(x_i) - \overline{R(x)}\right) \bullet \left(R(y_i) - \overline{R(y)}\right)\right)}{\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(R(x_i) - \overline{R(x)}\right)^2\right) \bullet \left(\frac{1}{n}\sum_{i=1}^{n}\left(R(y_i) - \overline{R(y)}\right)^2\right)}}$$

with x_1, x_2, x_n *depicting the predicted values;* y_1, y_2, y_n *depicting the real values and* n *representing the total number of predictions*

Time lag: time lag is a measure of the time shift between the actual and predicted glucose profile which results in the highest cross correlation coefficient between them[10,16]:

$$\tau_{delay} = \left(\check{y_k}(k - PH) * y(k)\right)$$

with y depicting the real values; $\hat{y}_k$ depicting the predicted values and PH depicting the prediction horizon.

Tables

S1 Table. Hyperparameter combinations evaluated in current experiments.

Hyperparameter	Values considered
Data preprocessing	
Normalization to [0, 1]	On, off
Back-propagation window	15, 30, 60, 120 minutes
Neural Network architecture	
RNN cell type	LSTM, RNN, GRU
RNN cell type, bi-directional structure	On, off
RNN number of hidden layers	1, 2, 3
RNN cell size	128, 64, 32, 16, 8, 4
RNN, activation function	ReLU, leaky ReLU, tanh, sigmoid, ELU
Dropout, presence	On, off
Dropout	0.05, 0.1, 0.2, 0.25
Model training	
Learning rate	1^{e}-2, 1^{e}-3, 1^{e}-4, 1^{e}-5
Learning rate scheduling	On, off
Learning rate scheduling decay	0.5, 0.6, 0.7, 0.8, 0.9, 0.95
Batch size	64, 128, 256, 512, 1024, 2048, 4096
Optimizer scheme	Adam, NAdam, RMSprop, SGD

S2 Table. Final set of hyperparameters for each of the machine learning models.

Hyperparameter	CGM-based glucose prediction	Combined glucose prediction
Data preprocessing		
Normalization to [0, 1]	On	On
Back-propagation window	30 minutes	30 minutes
Neural Network architecture		
RNN cell type	LSTM, LSTM, Dense	LSTM, LSTM, Dense
RNN cell type, bi-directional structure	Off, off, off	On, off, off
RNN number of hidden layers	3	3
RNN cell size	32, 16, 8	64, 32, 8
RNN, activation function	ReLU, ReLU, ReLU	ReLU, ReLU, ReLU
Dropout, presence	Off, off, on	Off, off, on
Dropout	0.1	0.1
Model training		
Learning rate	0.001	0.001
Learning rate scheduling	On	On
Learning rate scheduling decay	0.005 every 1,000 steps	0.005 every 1,000 steps
Batch size	1024	1024
Optimizer scheme	Adam	Adam

S3 Table. Extended baseline characteristics

Baseline characteristics for overall study population and subgroups normal glucose metabolism NGM) and prediabetes (PreD).

	CGM-based glucose prediction			**Combined glucose prediction**		
Characteristic	Total (n=851)	NGM (n=470)	PreD (n=184)	Total (n=540)	NGM (n=373)	PreD (n=99)
Age, years	59.9 ± 8.7	58.2 ± 8.8	61.5 ± 8.1	59.1 ± 8.7	58.1 ± 8.9	60.7 ± 8.5
Women, n (%)	418 (49.1)	266 (56.6)	83 (45.1)	276 (51.1)	207 (55.5)	47 (47.5)
BMI, kg/m^2	27.2 ± 4.4	25.6 ± 3.6	28.5 ± 4.4	26.5 ± 4.0	25.6 ± 3.5	28.8 ± 4.4
Newly diagnosed T2D, n (%)	70 (8.2)	-	-	35 (6.5)	-	-
Fasting plasma glucose, mmol/L	5.4 [5.0 – 6.2]	5.1 [4.8 – 5.4]	6.0 [5.4 – 6.3]	5.3 [4.9 – 5.8]	5.1 [4.8 – 5.4]	5.9 [5.3 – 6.2]
2-h post-load glucose, mmol/L	6.7 [5.2 – 9.1]*	5.5 [4.7 – 6.4]	8.4 [7.5 – 9.2]	6.2 [5.0 – 7.7]	5.5 [4.7 – 6.4]	8.4 [7.8 – 9.3]
HbA_{1c}, %	5.7 ± 0.8	5.4 ± 0.3	5.6 ± 0.4	5.6 ± 0.6	5.4 ± 0.3	5.5 ± 0.4
HbA_{1c}, mmol/mol	39.1 ± 8.3	35.4 ± 3.4	37.8 ± 4.2	37.3 ± 6.2	35.5 ± 3.4	37.1 ± 4.4
Sensor glucose Mean, mmol/L	6.1 [5.7 – 6.7]	5.8 [5.5 – 6.1]	6.2 [5.8 – 6.6]	5.9 [5.6 – 6.4]	5.9 [5.5 – 6.1]	6.2 [5.7 – 6.6]
SD, mmol/L	0.84 [0.68 – 1.18]	0.73 [0.62 – 0.87]	0.90 [0.74 – 1.13]	0.79 [0.66 – 1.01]	0.73 [0.63 – 0.89]	0.89 [0.73 – 1.11]
SD > 1.37 mmol/L, n (%)	142 (16.7)	11 (2.3)	16 (8.7)	50 (9.3)	9 (2.4)	5 (5.1)
CV, %	14.0 [11.6 – 17.6]	12.6 [10.8 – 14.9]	14.9 [12.2 – 17.5]	13.3 [11.2 – 16.8]	12.8 [10.9 – 15.2]	14.7 [12.1 – 17.5]
Diabetes medication use, n (%)	109 (12.8)†	-	-	27 (4.8)	-	-
Insulin	19 (2.2)	-	-	4 (0.7)	-	-
Metformin	104 (12.2)	-	-	27 (5.0)	-	-
Sulfonylureas	21 (2.5)	-	-	6 (1.1)	-	-
Thiazolidinediones	0 (0)	-	-	0 (0)	-	-
GLP-1 analogs	3 (0.4)	-	-	1 (0.2)	-	-
DDP-4 inhibitors	1 (0.1)	-	-	0 (0)	-	-
SGLT-2 inhibitors	1 (0.1)	-	-	0 (0)	-	-

	CGM-based glucose prediction			Combined glucose prediction		
Office SBP, mmHg	133.3 ± 18.0	129.3 ± 17.5	137.2 ± 19.2	132.2 ± 17.9	129.6 ± 17.6	138.1 ± 18.6
Office DBP, mmHg	75.2 ± 10.2	73.5 ± 9.8	76.7 ± 10.3	74.7 ± 10.1	73.5 ± 9.8	77.2 ± 10.8
Antihypertensive medication use, n (%)	305 (35.9)†	105 (22.3)	74 (40.2)	162 (30.0)	84 (22.5)	37 (37.4)
Total-to-HDL cholesterol ratio	3.5 [2.8 – 4.3]	3.3 [2.8 – 4.4]	3.8 [3.1 – 4.7]	3.4 [2.8 – 4.3]	3.3 [2.8 – 4.3]	3.6 [2.9 – 4.5]
Triglycerides, mmol/L	1.3 [0.9 – 1.8]	1.1 [0.8 – 1.5]	1.4 [1.0 – 2.0]	1.2 [0.9 – 1.7]	1.1 [0.8 – 1.5]	1.4 [1.0 – 1.9]
Lipid-modifying medication use, n (%)	212 (24.9)†	52 (11.1)	45 (24.5)	100 (18.5)	38 (10.2)	23 (23.2)
Smoking status						
Never/former/current, n	327/415/106‡	198/210/60	62/101/20	214/253/70	160/164/47	35/52/10
Never/former/current, %	38.6/48.9/12.5	42.3/44.9/12.8	33.9/55.2/10.9	39.9/47.1/13.0	43.1/44.2/12.7	35.7/54.1/10.2

*Data are reported as mean ± SD, median [interquartile range], or number (percentage [%]) as appropriate. CGM, continuous glucose monitoring; NGM, normal glucose metabolism; PreD, prediabetes; BMI, body mass index; HbA1c, glycated hemoglobin A1c; S D, standard deviation; CV, coefficient of variation; GLP-1, glucagon-like peptide-1; DPP-4, dipeptidase-4; SGLT-2, sodium-glucose cotransporter 2; SBP, systolic blood pressure; DBP, diastolic blood pressure; HDL, high-density lipoprotein. * Missing in 38 participants; † missing in one participant; ‡ missing in three participants.*

S4 Table. Extended analysis of model performance in normal glucose metabolism and prediabetes subgroups.

		CGM-based glucose prediction		Combined glucose prediction	
		NGM (n=92)	**PreD (n=35)**	**NGM (n=75)**	**PreD (n=21)**
15 minutes	RMSE, mmol/L	0.158 [0.154 – 0.161]	0.238 [0.233 – 0.343]	0.151 [0.149 – 0.153]	0.232 [0.227 – 0.237]
	< 5% , %	93.56 [93.50 - 93.62]	92.58 [92.51 – 92.66]	93.76 [93.71 – 93.80]	92.65 [92.57 – 97.73]
	< 10% , %	99.47 [99.42 – 99.52]	99.01 [98.98 – 99.05]	99.65 [99.62 – 99.68]	99.08 [99.04 – 99.12]
	Rho	0.951 [0.948 – 0.954]	0.973 [0.967 – 0.980]	0.953 [0.950 – 0.957]	0.974 [0.968 – 0.979]
60 minutes	RMSE, mmol/L	0.501 [0.498 – 0.505]	0.602 [0.594 – 0.610]	0.503 [0.495 – 0.510]	0.599 [0.594 – 0.604]
	< 5% , %	74.19 [74.11 – 74.26]	69.48 [69.39 – 69.56]	75.02 [74.95 – 75.08]	70.01 [69.95 – 70.07]
	< 10% , %	89.89 [89.82 – 89.97]	87.43 [87.36 – 87.50]	89.25 [89.19 – 89.31]	88.20 [88.09 – 88.30]
	Rho	0.699 [0.697 – 0.702]	0.732 [0.727 – 0.738]	0.701 [0.697 – 0.705]	0.739 [0.732 – 0.747]

Data are reported as mean [95% confidence interval]. CGM, continuous glucose monitoring; NGM, normal glucose metabolism; PreD, prediabetes; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

S5 Table. Extended analysis on time lag between predicted and actual glucose values.

Time lag		Total (n=170)	NGM (n=92)	PreD (n=35)	T2D (n=43)
15 minutes	CGM-based	0.12 ± 0.18	0.08 ± 0.12	0.07 ± 0.11	0.41 ± 0.92
	Combined	0.17 ± 0.11	0.11 ± 0.18	0.10 ± 0.18	0.44 ± 0.38
60 minutes	CGM-based	12.28 ± 6.84	7.02 ± 3.46	9.03 ± 3.77	14.92 ± 11.18
	Combined	11.95 ± 7.32	7.19 ± 2.87	8.75 ± 3.91	14.28 ± 9.96

Data are reported as mean ± SD. NGM, normal glucose metabolism; PreD, prediabetes; T2D, type 2 diabetes.

Time lag		Total (n=170)	NGM (n=92)	PreD (n=35)	T2D (n=43)
15 minutes	CGM-based	0 [0 – 5]	0 [0 – 0]	0 [0 – 0]	0 [0 – 0]
	Combined	0.50 [0.25 – 0.75]	0.25 [0.0 – 0.75]	0.25 [0 – 0.50]	0.50 [0.25 – 0.75]
60 minutes	CGM-based	10 [5 – 15]	10 [0 – 15]	10 [5 – 15]	15 [5 – 20]
	Combined	9.50 [4.25 – 16.50]	6.75 [4.25 – 9.50]	7.50 [4.50 – 10.75]	14.50 [6.75 – 21.50]

Data are reported as median [IQR]. NGM, normal glucose metabolism; PreD, prediabetes; T2D, type 2 diabetes.

S6 Table. Extended analysis of model performance with t_0 glucose value as predictor.

Prediction with t_0		Total (n=170)	NGM (n=92)	PreD (n=35)	T2D (n=43)
15 minutes	RMSE, mmol/L	0.158 [0.154 – 0.161]	0.238 [0.233 – 0.343]	0.151 [0.149 – 0.153]	0.232 [0.227 – 0.237]
	< 5% , %	93.56 [93.50 - 93.62]	92.58 [92.51 – 92.66]	93.76 [93.71 – 93.80]	92.65 [92.57 – 97.73]
	< 10% , %	99.47 [99.42 – 99.52]	99.01 [98.98 – 99.05]	99.65 [99.62 – 99.68]	99.08 [99.04 – 99.12]
	Rho	0.951 [0.948 – 0.954]	0.973 [0.967 – 0.980]	0.953 [0.950 – 0.957]	0.974 [0.968 – 0.979]
60 minutes	RMSE, mmol/L	0.501 [0.498 – 0.505]	0.602 [0.594 – 0.610]	0.503 [0.495 – 0.510]	0.599 [0.594 – 0.604]
	< 5% , %	74.19 [74.11 – 74.26]	69.48 [69.39 – 69.56]	75.02 [74.95 – 75.08]	70.01 [69.95 – 70.07]
	< 10% , %	89.89 [89.82 – 89.97]	87.43 [87.36 – 87.50]	89.25 [89.19 – 89.31]	88.20 [88.09 – 88.30]
	Rho	0.699 [0.697 – 0.702]	0.732 [0.727 – 0.738]	0.701 [0.697 – 0.705]	0.739 [0.732 – 0.747]

Data are reported as mean [95% confidence interval]. NGM, normal glucose metabolism; PreD, prediabetes; T2D, type 2 diabetes; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

S7 Table. Model performance stratified by day and night.

15 minutes		CGM-based glucose prediction		Combined glucose prediction	
		Total (n=170)	T2D (n=43)	Total (n=109)	T2D (n=13)
Day	RMSE, mmol/L	0.199 [0.196 – 0.202]	0.300 [0.295 – 0.305]	0.197 [0.193 – 0.201]	0.287 [0.283 – 0.291]
	< 5% , %	92.92 [92.85 – 92.99]	91.97 [91.87 – 92.07]	92.94 [92.90 – 92.98]	91.95 [91.92 – 91.98]
	< 10% , %	99.11 [99.07 – 99.15]	98.84 [98.72 – 98.95]	99.17 [99.13 – 99.21]	98.82 [98.78 – 98.86]
	Rho	0.955 [0.952 – 0.958]	0.984 [0.981 – 0.987]	0.964 [0.962 – 0.966]	0.986 [0.984 – 0.988]
Night	RMSE, mmol/L	0.182 [0.175 – 0.189]	0.278 [0.273 – 0.283]	0.178 [0.173 – 0.183]	0.257 [0.252 – 0.262]
	< 5% , %	93.08 [93.01 – 93.15]	92.07 [91.99 – 92.15]	93.11 [93.07 – 93.15]	92.14 [92.10 – 92.18]
	< 10% , %	99.28 [99.20 – 99.36]	98.94 [98.88 – 99.00]	99.34 [99.30 – 99.38]	99.03 [98.98 – 99.08]
	Rho	0.967 [0.964 – 0.970]	0.989 [0.986 – 0.992]	0.974 [0.970 – 0.978]	0.994 [0.992 – 0.996]

60 minutes		CGM-based glucose prediction		Combined glucose prediction	
		Total (n=170)	T2D (n=43)	Total (n=109)	T2D (n=13)
Day	RMSE, mmol/L	0.687 [0.683 – 0.691]	0.775 [0.768 – 0.783]	0.536 [0.532 – 0.540]	0.768 [0.760 – 0.776]
	< 5% , %	68.40 [68.35 – 68.45]	63.69 [63.57 – 63.81]	68.87 [68.80 – 68.94]	64.03 [63.95 – 64.11]
	< 10% , %	85.27 [85.20 – 85.35]	83.82 [83.77 – 83.87]	86.12 [86.07 – 86.17]	84.08 [84.01 – 84.15]
	Rho	0.640 [0.629 – 0.651]	0.703 [0.697 – 0.709]	0.658 [0.654 – 0.663]	0.710 [0.705 – 0.715]
Night	RMSE, mmol/L	0.497 [0.491 – 0.503]	0.634 [0.629 – 0.639]	0.512 [0.503 – 0.521]	0.633 [0.627 – 0.639]
	< 5% , %	72.61 [72.56 – 72.66]	69.42 [69.33 – 69.51]	71.58 [71.47 – 71.69]	69.28 [69.21 – 69.35]
	< 10% , %	89.44 [89.31 – 89.57]	87.10 [87.01 – 87.20]	88.78 [88.68 – 88.88]	87.30 [87.19 – 87.41]
	Rho	0.793 [0.784 – 0.802]	0.854 [0.848 – 0.860]	0.783 [0.780 – 0.786]	0.861 [0.855 – 0.867]

Data are reported as mean [95% confidence interval]. CGM, continuous glucose monitoring; T2D, type 2 diabetes; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

S8 Table. Model performance stratified by low versus high glucose variability.

		CGM-based glucose prediction		Combined glucose prediction	
		SD ≤ 1.37 mmol/L (n=142)	**SD > 1.37 mmol/L (n=28)**	**SD ≤ 1.37 mmol/L (n=101)**	**SD > 1.37 mmol/L (n=8)**
15 minutes	RMSE, mmol/L	0.179 [0.176 – 0.181]	0.301 [0.289 – 0.313]	0.180 [0.177 – 0.183]	0.288 [0.276 – 0.299]
	< 5% , %	93.02 [92.99 – 93.05]	91.88 [91.77 – 92.05]	93.21 [93.18 – 93.24]	92.00 [91.92 – 92.08]
	< 10% , %	99.25 [99.22 – 99.28]	98.79 [98.76 – 98.84]	99.30 [99.29 – 99.32]	98.82 [98.73 – 98.91]
	Rho	0.960 [0.959 – 0.962]	0.983 [0.980 – 0.986]	0.965 [0.964 – 0.967]	0.992 [0.988 – 0.996]
60 minutes	RMSE, mmol/L	0.549 [0.542 – 0.555]	0.711 [0.699 – 0.724]	0.559 [0.552 – 0.565]	0.710 [0.700 – 0.719]
	< 5% , %	71.04 [70.89 – 71.21]	65.33 [65.19 – 66.46]	71.81 [71.77 – 71.85]	66.17 [66.09 – 66.23]
	< 10% , %	89.19 [89.15 – 89.22]	84.42 [84.28 – 84.56]	90.01 [89.95 – 90.06]	85.25 [85.11 – 85.39]
	Rho	0.701 [0.700 – 0.702]	0.741 [0.737 – 0.745]	0.723 [0.719 – 0.728]	0.801 [0.784 – 0.817]

Data are reported as mean [95% confidence interval]. CGM, continuous glucose monitoring; SD, standard deviation; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

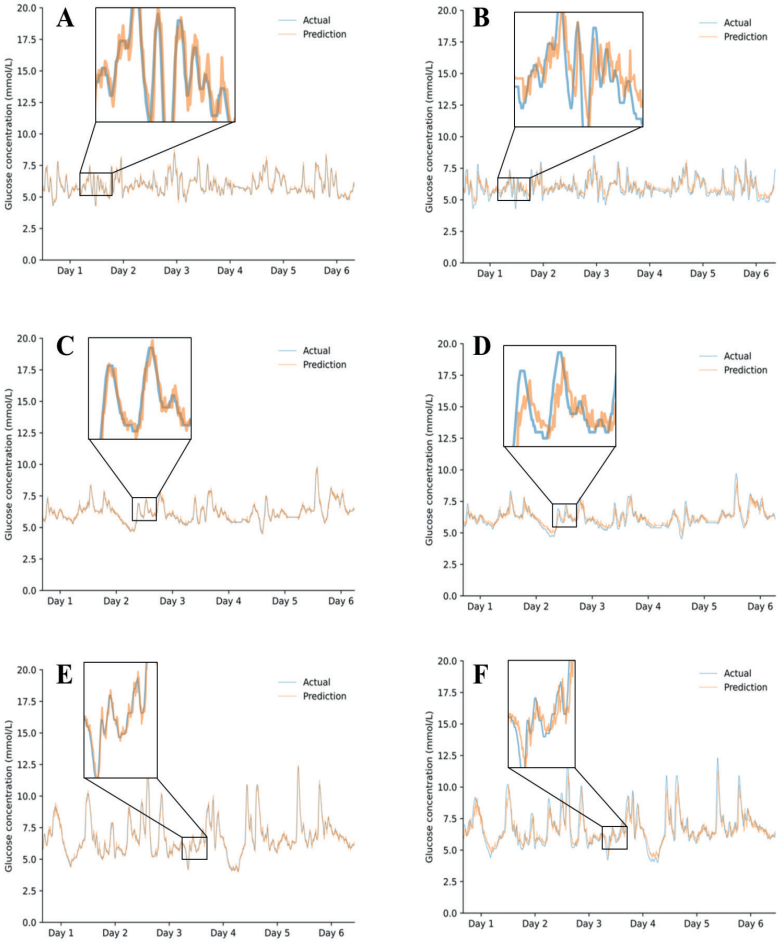
S9 Table. Extended analysis of model performance in the OhioT1DM Dataset

		15 minutes	**30 minutes**	**60 minutes**
Main model	RMSE, mmol/L	0.689 [0.685 – 0.693]	1.189 [1.183 – 1.195]	1.918 [1.910 – 1.926]
	< 5% , %	61.84 [61.64 – 61.04]	39.10 [38.86 – 39.34]	22.28 [22.01 – 22.65]
	< 10% , %	86.02 [85.83 – 86.21]	64.72 [64.45 – 64.99]	40.19 [39.89 – 40.50]
	Rho	0.908 [0.905 – 0.911]	0.792 [0.789 – 0.795]	0.605 [0.602 – 0.608]
Optimized model	RMSE, mmol/L	0.426 [0.422 – 0.430]	1.046 [1.039 – 1.052]	1.733 [1.725 – 1.741]
	< 5% , %	72.22 [71.95 – 72.47]	48.99 [48.77 – 49.22]	39.85 [39.55 – 40.15]
	< 10% , %	91.22 [91.01 – 91.43]	71.48 [71.24 – 71.73]	49.81 [49.48 – 50.14]
	Rho	0.948 [0.946 – 0.950]	0.886 [0.884 – 0.888]	0.689 [0.686 – 0.692]

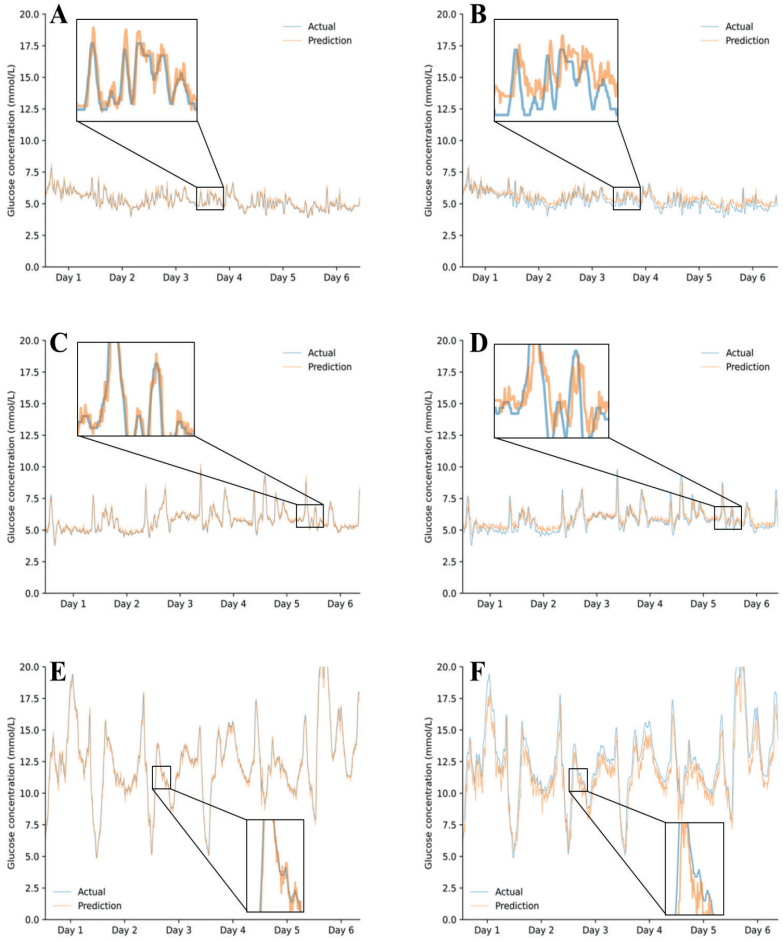
Data are reported as mean [95% confidence interval]. Main model: CGM-based model trained on main study population; Optimized model; main CGM-based model trained on main study population and portion of data from individuals with type 1 diabetes; RMSE, root-mean-square error; < 5%, percentage of predicted values within 5% of actual glucose values; < 10%, percentage of predicted values within 10% of actual glucose values; rho, Spearman's rank correlation coefficient.

Figures

S1 Figure. Illustrative examples of continuous glucose monitoring-based machine learning model predictions compared to actual values.
Predictions in an individual with normal glucose metabolism (NGM) on a time interval of 15 (A) and 60 (B) minutes. Predictions in an individual with prediabetes (PreD) on a time interval of 15 (C) and 60 (D) minutes. Predictions in an individual with type 2 diabetes (T2D) on a time interval of 15 (E) and 60 (F) minutes.

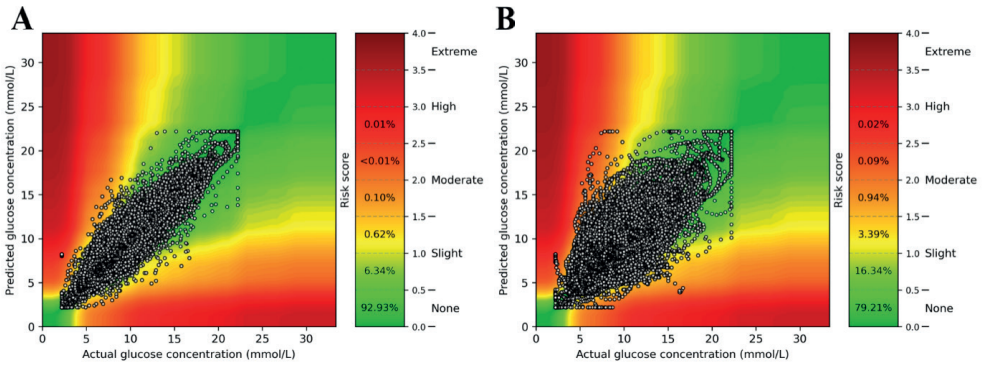


S2 Figure. Illustrative examples of continuous glucose monitoring- and accelerometry-based machine learning model predictions compared to actual values.
Predictions in an individual with normal glucose metabolism (NGM) on a time interval of 15 (A) and 60 (B) minutes. Predictions in an individual with prediabetes (PreD) on a time interval of 15 (C) and 60 (D) minutes. Predictions in an individual with type 2 diabetes (T2D) on a time interval of 15 (E) and 60 (F) minutes.



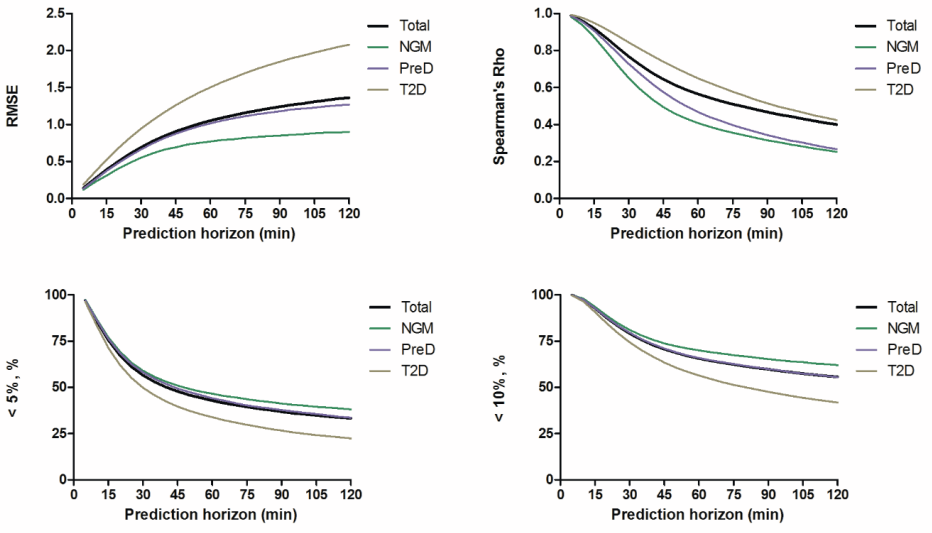
S3 Figure. Surveillance error grid evaluation of glucose prediction safety at time intervals of 15 and 60 minutes using glucose value t0 as predictor.

Assessment of glucose prediction safety in individuals with type 2 diabetes (n=43) at 15 minutes (panel A) and 60 minutes (panel B) using a naïve approach with t0 as predictor. The risk score values translate to the following degrees of risk: 0 - 0.5, none; 0.5 - 1.0, slight (lower); 1.0 - 1.5, slight (higher); 1.5 - 2.0, moderate (lower); 2.0 - 2.5, moderate (higher); 2.5 – 3.0, great (lower); 3.0 - 3.5, great (higher); > 3.5 extreme.

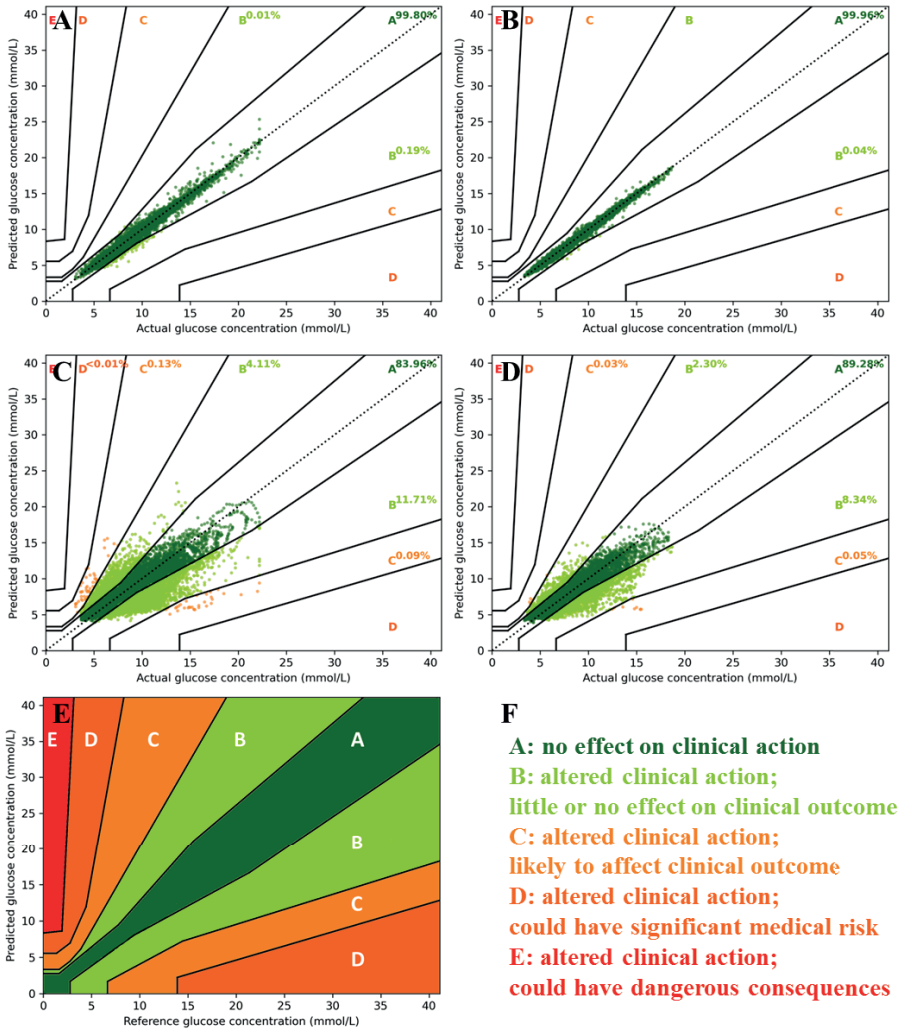


S4 Figure. Performance characteristics of a prediction model using t0 as predictor across time horizons between 0 and 120 minutes.

An extended analysis of model performance using t0 glucose value as predictor was carried out for normal glucose metabolism (NGM), prediabetes (PreD) and individuals with type 2 diabetes (T2D). Models were evaluated using RMSE (root-mean-square error; upper-left), Spearman's rank correlation coefficient (upper-right) and percentage of predicted values within 5% and 10% of actual glucose values, respectively (lower-left and right).

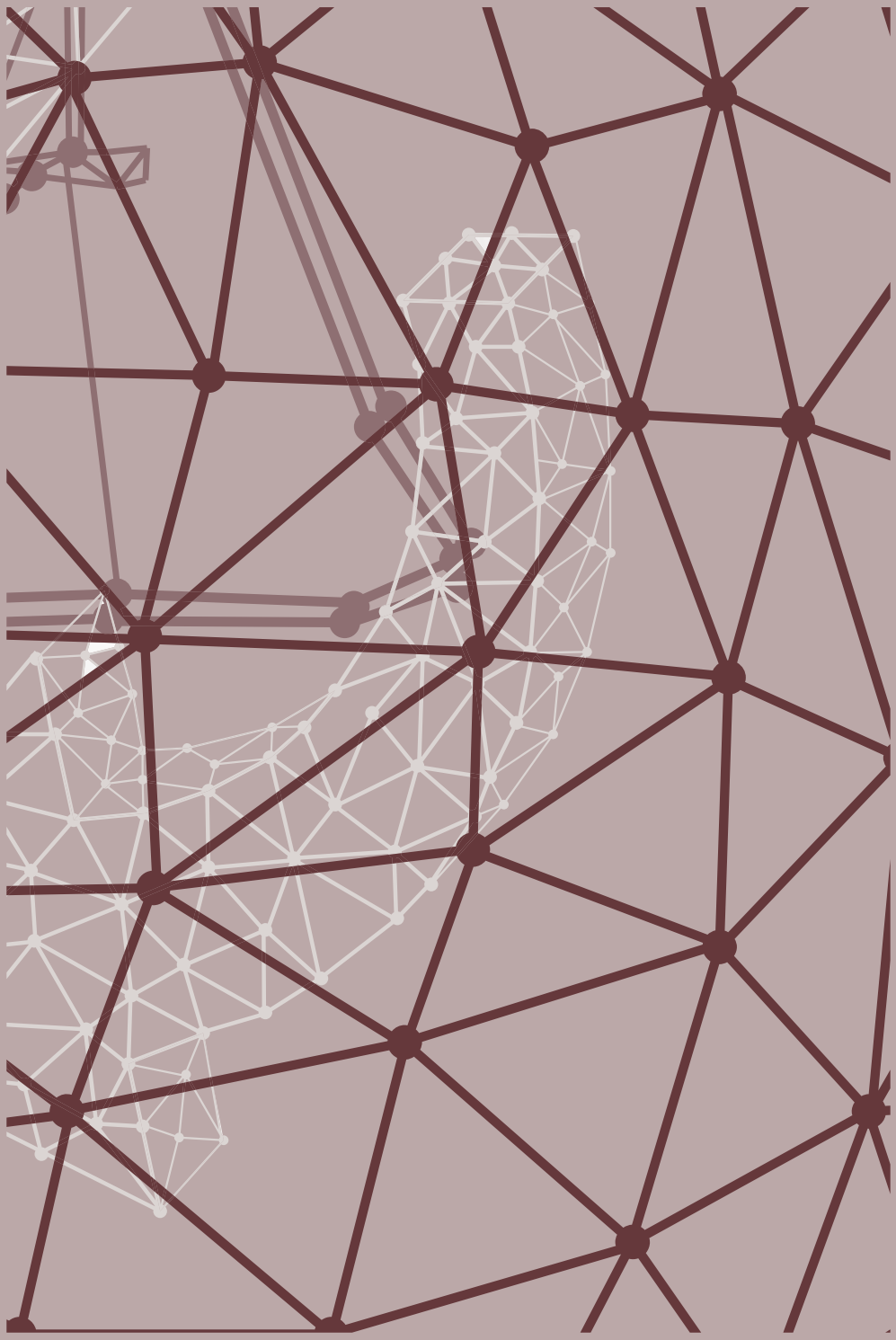


S5 Figure. Parkes error grid evaluation of glucose prediction safety at time intervals of 15 and 60 minutes. Assessment of CGM-based glucose prediction safety in individuals with type 2 diabetes (n=43) at 15 minutes (panel A) and 60 minutes (panel C). Assessment of CGM- and accelerometry-based glucose prediction safety in individuals with type 2 diabetes (n=13) at 15 minutes (panel B) and 60 minutes (panel D). Each error zone (panel E) has a different clinical interpretation and consequence (panel F). A horizontal shift towards zone E represents overestimation by the algorithm (higher predicted than actual glucose values); a vertical shift towards zone D represents underestimation (lower predicted than actual glucose values).



Supplemental references

1. Otoom M, Alshraideh H, Almasaeid HM, Lopez-de-Ipina D, Bravo J. Real-Time Statistical Modeling of Blood Sugar. J Med Syst. 2015;39(10):123. Epub 2015/08/26. doi: 10.1007/s10916-015-0301-8. PubMed PMID: 26303151.
2. Yang J, Li L, Shi Y, Xie X. An ARIMA Model With Adaptive Orders for Predicting Blood Glucose Concentrations and Hypoglycemia. IEEE J Biomed Health Inform. 2019;23(3):1251-60. Epub 2018/07/12. doi: 10.1109/JBHI.2018.2840690. PubMed PMID: 29993728.
3. Sun Q, Jankovic MV, Bally L, Mougiakakou SG. Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network. arXiv e-prints [Internet]. 2018 September 01, 2018. Available from: https://ui.adsabs.harvard.edu/abs/2018arXiv180903817S.
4. Rodriguez-Rodriguez I, Chatzigiannakis I, Rodriguez JV, Maranghi M, Gentili M, Zamora-Izquierdo MA. Utility of Big Data in Predicting Short-Term Blood Glucose Levels in Type 1 Diabetes Mellitus Through Machine Learning Techniques. Sensors (Basel). 2019;19(20). Epub 2019/10/19. doi: 10.3390/s19204482. PubMed PMID: 31623111; PubMed Central PMCID: PMCPMC6833040.
5. Georga EI, Protopappas VC, Ardigo D, Polyzos D, Fotiadis DI. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions. Diabetes Technol Ther. 2013;15(8):634-43. Epub 2013/07/16. doi: 10.1089/dia.2012.0285. PubMed PMID: 23848178.
6. Hamdi T, Ben Ali J, Di Costanzo V, Fnaiech F, Moreau E, Ginoux J-M. Accurate prediction of continuous blood glucose based on support vector regression and differential evolution algorithm. Biocybernetics and Biomedical Engineering. 2018;38(2):362-72. doi: https://doi.org/10.1016/j.bbe.2018.02.005.
7. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017:3146--54.
8. Allam F, Nossair Z, Gomma H, Ibrahim I, Salam MA-e, editors. Prediction of subcutaneous glucose concentration for type-1 diabetic patients using a feed forward neural network. The 2011 International Conference on Computer Engineering & Systems; 2011 29 Nov.-1 Dec. 2011.
9. Zarkogianni K, Mitsis K, Litsa E, Arredondo MT, Ficomicron G, Fioravanti A, et al. Comparative assessment of glucose prediction models for patients with type 1 diabetes mellitus applying sensors for glucose and physical activity monitoring. Med Biol Eng Comput. 2015;53(12):1333-43. Epub 2015/06/08. doi: 10.1007/s11517-015-1320-PubMed PMID: 26049412.
10. Perez-Gandia C, Facchinetti A, Sparacino G, Cobelli C, Gomez EJ, Rigla M, et al. Artificial neural network algorithm for online glucose prediction from continuous glucose monitoring. Diabetes Technol Ther. 2010;12(1):81-8. Epub 2010/01/20. doi: 10.1089/dia.2009.0076. PubMed PMID: 20082589.
11. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9(8):1735-80. doi: 10.1162/neco.1997.9.8.1735.
12. Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. arXiv e-prints. 2018:arXiv:1808.03314.
13. Staudemeyer RC, Rothstein Morris E. Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv e-prints [Internet]. 2019 September 01, 2019. Available from: https://ui.adsabs.harvard.edu/abs/2019arXiv190909586S.
14. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arXiv e-prints. 2014:arXiv:1412.3555.
15. Schuster M, Paliwal K. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on. 1997;45:2673-81. doi: 10.1109/78.650093.
16. Li K, Liu C, Zhu T, Herrero P, Georgiou P. GluNet: A Deep Learning Framework for Accurate Glucose Forecasting. IEEE J Biomed Health Inform. 2020;24(2):414-23. Epub 2019/08/02. doi: 10.1109/JBHI.2019.2931842. PubMed PMID: 31369390.



CHAPTER 11

DISCUSSION

Introduction

There is a continuous strive for faster, more accurate and overall better biomarkers in daily clinical practice. Despite large investments and monumental technological advances in biomarker discovery, it is estimated that on average less than one diagnostic protein biomarker is approved by the US Food and Drug Administration (FDA) on a yearly basis[1]. Thus, there is a large gap between the number of new biomarkers being studied versus the actual translation of biomarkers from discovery phase to clinical practice. This thesis built on an alternative strategy to improve the diagnostic trajectory or prognostication of patients by optimizing current biomarkers, rather than deploying new ones.

The first part of this thesis reviewed the analytical and clinical performance of cardiac troponin and N-terminal prohormone of brain natriuretic peptide (NT-proBNP) according to the biomarker evaluation framework. Clinical laboratory practice recommendations and differences between Europa and the USA were reviewed (Chapter 2), the influence of biotin on the analytical performance was assessed (Chapter 3) and the kinetics after myocardial infarction were studied (Chapter 4) for cardiac troponin. Additionally, the impact of a potential circadian rhythm on the clinical performance of NT-proBNP was evaluated (Chapter 5).

The second part of this thesis focused on optimizing current routine biomarkers. In Chapter 6, a novel concept detecting 2-nitroimidazole modifications on troponin was demonstrated as an attempt to improve the diagnostic specificity for myocardial infarction. In chapters 7 to 10, the development of clinical decision support systems that interpretate multiple biomarker results were studied for the emergency department (Chapters 7, 8, 9) and for continuous glucose prediction in individuals with diabetes (Chapter 10). This discussion section appraises the major findings of this thesis and provides directions towards further research.

<u>The interpretation and global implementation of cardiac troponin testing</u>

Chapters 2 and 4 assessed the interpretation and implementation of cardiac troponin testing. The cardiac troponin complex consists of the proteins troponin T, I and C with cardiac troponin T (cTnT) and I (cTnI) expressed exclusively in the heart and therefore considered cardiac specific biomarkers for myocardial infarction (MI)[2]. Despite their biochemical differences, cTnT and cTnI have equally high diagnostic accuracy for MI and are therefore considered interchangeably[3,4]. The current clinical guidelines for MI recommend serial troponin assessment, and it is therefore important to understand the kinetics of cardiac troponins in blood after MI. Chapter 4 demonstrated that there are substantial differences in the kinetics after MI, especially in later stages, with cTnT exhibiting a bi-phasic curve and cTnI showing a monophasic curve. These findings were recently confirmed in literature[5,6], enlarging the body of evidence that demonstrates

substantial differences between troponin T and troponin I kinetics[7]. Importantly, these findings may suggest that result interpretation in very late presenters (> 20 hours) could be different for troponin T vs troponin I.

The interpretation of cardiac troponin results is guided by various clinical guidelines. Naturally, standardization and harmonization of these guidelines is therefore critical to ensure consistent, high-quality clinical care across clinical laboratories. Chapter 2 provided recommendations for laboratories implementing cardiac troponin testing, and reviewed existing differences between Europe and the United States (US). These differences were further appraised in recent studies[8-11], which examined clinical protocols, sampling times, decision limits and test utilization. Researchers found substantial heterogeneity in cardiac troponin and natriuretic peptide testing across institutes[9,12], with for example large differences in derivation of decision limits for cardiac troponin; seven different approaches to decision limits were used varying between the 10% CV (16.2% of the laboratories), 20% CV (2.0%), the 99th reference interval (52.3%) or even local clinical decisions (44%). These studies prove that there is an important gap between guideline established recommendations and actual real-world implementation. It is important to notice that several clinical guidelines for cardiac troponin utilization are applicable[13], as most of them are very similar in terms of rule-out performance. Hence, it is highly recommended that institutes implement at least one of these guidelines, with their choice informed by local priorities and the multidisciplinary team.

Increasing the clinical specificity of cardiac troponin by detection of 2-nitroimidazole modifications

The introduction of the latest high-sensitive immunoassay for cardiac troponin (hs-cTn) allowed even more accurate measurements in the lower analytical range. This enabled clinicians to diagnose myocardial infarction (MI) even faster and substantially reduced the proportion of patients that were classified with unstable angina pectoris (UAP). As a consequence of these more accurate measurements in the lower analytical range, it became apparent that troponin levels were detected in pathological conditions other than MI. In fact, troponin levels were detected even in non-cardiac conditions and in presumably healthy individuals. These developments are illustrated by the clinical performance of the high-sensitive immunoassay, with high sensitivity and negative predictive value (>99%) but suboptimal specificity and positive predictive value (±75%). Hence, an unmet clinical need has become a cardiac biomarker assay that can discriminate between hypoxia-induced cardiomyocyte damage (AMI) and cardiomyocyte damage other than AMI. In Chapter 6 an innovative concept was introduced allowing the selective detection of cardiac troponin released under conditions of hypoxia, but leaving biomarker release due to non-hypoxic injury triggers undetected. In particular, a family of compounds called 2-nitroimidazole were employed, which have successfully been applied in imaging hypoxic tumors in the

past decades. These moieties were demonstrated to selectively modify proteins under conditions of hypoxia, leading to the development of a novel mass spectrometry assay that detects 2-nitroimidazole modifications on proteins (such as troponin). This concept potentially addressed the unmet clinical need of detecting hypoxia-induced cardiomyocyte damage, but future research is now warranted to examine the feasibility of this method in cellular systems of hypoxia. Ultimately, this concept should be evaluated in patients with chest pain presenting to the cardiac emergency department.

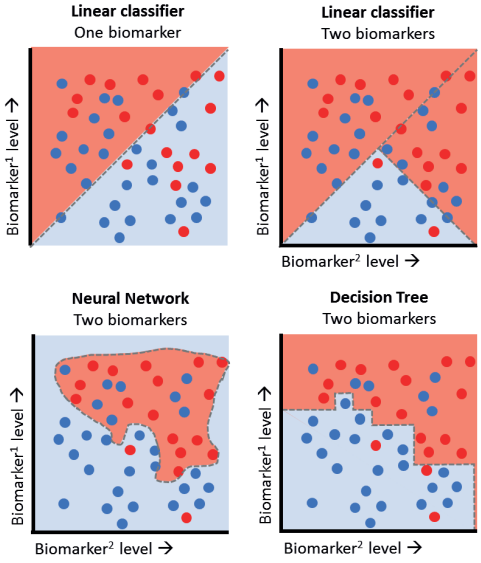
Apart from the 2-nitroimidazole detection method, other approaches enhancing the specificity of cardiac troponins have been described in the literature[7,14,15]. The most promising of these include the combination of troponin with a second (often ischemic) biomarker (such as cardiac-myosin binding protein C [cMyc], ischemia-modified albumin [IMA], and heart-fatty acid binding protein [H-FABP]) and the detection of specific degraded molecular fragments of cardiac troponin in the blood[7,16-18]. Multiple studies found that molecular troponin fragments after MI are different compared to conditions including end-stage renal disease[19] and physical exercise[20]. Due to a variety of reasons including insufficient performance, technical difficulties and associated costs, each of these approaches have not yet succeeded in enhancing the specificity of cardiac troponins.

Combining biomarker levels information using machine learning into interpretable clinical outputs

The first chapters of this thesis examined routine biomarkers cardiac troponin and NT-proBNP. Both biomarkers were applied in a specific clinical context, such as myocardial infarction or acute heart failure diagnostics. This approach is limited by the fact that information from only one biomarker is used. It is therefore of particular interest to move a step further by developing clinical decision support systems that interpretate multiple biomarker results into a more readily interpretable output for a clinical user, such as mortality risk or likelihood of myocardial infarction, leading to more information being used, potentially enhancing the performance for a specific clinical task.

The combination of large amounts of biomarker data is reasonably hard for humans, especially when the number of dimensions exceed three. This is where statistical and machine learning techniques come into play. These techniques can be categorized from simple, e.g. linear classifiers and support-vector machines, to complicated, e.g. neural networks and high-dimensional decision trees. In general, the more sophisticated an algorithm is, the more complex patterns can be captured in the biomarker data (Figure 1). To date, even the most complex algorithms are now feasible and practical in routine clinical care, as large amounts of computing power and data are readily available.

Figure 1: Decision boundaries of four statistical and machine learning classifiers. Dots represent individual data points being either positive (red) or negative (blue). Shaded background represents the classification of the algorithm with the dotted line representing the decision boundary.



This thesis applied algorithms of varying complexity as a means to enhance the application of biomarkers in two clinical situations. First, a clinical decision support system was built for the risk stratification of patients in the emergency department (Chapters 7, 8, 9). Second, historical glucose and physical activity data were integrated from individuals with diabetes to predict their future glucose values (Chapter 10).

Combining biomarker data for the rapid risk stratification of patients in the emergency department

Risk stratification of patients presenting to the emergency department (ED) is important for appropriate triage and early treatment. Hence, a novel clinical decision support tool was developed predicting the 31-day mortality risk of patients after two hours by combining all available biomarker data (Chapters 7, 8). In a pilot study with a small group of sepsis patients, this tool outperformed established clinical risk scores and internal medicine physicians in predicting the mortality risk (Chapter 7). Subsequently, the tool was assessed in a large, multi-center cohort, demonstrating high diagnostic performance for predicting mortality with area under the receiver-operating-characteristic curve (ROC) ranging from 0.88 to 0.98 in four large hospitals in The Netherlands (Chapter 8). Additionally, the clinical decision support tool was designed to visualize patient characteristics and laboratory data patterns that underlie individual mortality predictions, thereby partly mitigating the issue of “black-box” predictions.

Even though the newly developed clinical decision support tool has high diagnostic performance in a retrospective setting, its actual real-world performance and true clinical value are unknown. Follow-up studies are carried out studying both the model performance and its clinical value. It is important to note that high performance in a retrospective setting does not automatically translate to a prospective, real-world setting. In fact, a growing body of recent evidence shows that a majority of promising, high-performance machine learning models in the development phase demonstrate fairly discouraging performance in a real-world setting, thereby also directly losing their additional clinical value[21,22]. Once again this highlights that proper evaluation in a prospective, real-world setting is absolutely the cornerstone in determining the actual value of machine-learning based clinical decision support tools.

Historical glucose data for the continuous prediction of future glucose values in individuals with diabetes

Closed-loop insulin delivery systems that integrate continuous glucose monitoring, insulin infusion and a control algorithm to continuously regulate blood glucose levels are increasingly being used to keep individuals with diabetes in a desirable glucose range. Certain obstacles to the optimal performance of these devices remain and prediction of future glucose values would aid in overcoming these obstacles. In Chapter 10, historical glucose and physical activity data were combined in healthy individuals and individuals with (pre-)diabetes to predict their future glucose values on a 15- and 60-minute interval. It was demonstrated that predictions were not only accurate in individuals with type 1 and 2 diabetes, but also clinically safe as demonstrated by surveillance error grids. These results hint at a potential role of such prediction models to overcome the ~10-minute sensor delay, which, in part, results from measuring interstitial glucose rather than blood glucose concentrations, and short periods of sensor malfunction which occasionally occur in closed-loop insulin dosing systems[23,24]. Moreover, the predictions on a 60-minute interval might aid in early anticipation on large glucose fluctuations.

To our knowledge, this was the first study to report this level of performance in a large, population based sample of individuals with type 1 and type 2 diabetes[25-28]. The high level of performance partly could be explained by the unique approach of including glucose and physical activity profiles from healthy individuals and individuals with prediabetes. Initially this might seem a bit contradictory, but as a consequence, the models first "learned" to recognize patterns in glucose profiles of these healthy individuals, followed by "adaptation" with glucose patterns of individuals with type 1 and 2 diabetes. Such approach -called transfer learning- not only allowed to employ a larger sample size, but most likely also enriched the generalizability and robustness of these models. Transfer learning has recently seen enormous success in tasks such as image classification[24,25], natural-language processing[26,27] and is now starting to become more commonly used

for clinical purposes, such as in radiology[28,29] and pathology[30,31]. An increased usage of transfer learning among medical predictions models will most likely lead to more robust and overall better models.

Concluding remarks

The laboratory measurement of a biomarker is cornerstone in the majority of day-to-day clinical decision making. This thesis examined multiple aspects of routine clinical biomarkers and their optimization to address unmet diagnostic challenges. In the first part, various characteristics of the biomarker evaluation framework were examined for cardiac troponin and NT-proBNP. In the second part, alternative approaches to improve the diagnostic trajectory or prognostication of patients by optimizing current biomarkers, rather than deploying new ones, were demonstrated. Based on our main findings, a number of future directions are proposed.

First, standardization of biomarker testing across institutes will likely lead to substantial health benefits, presumably often also larger than one would achieve by optimization of an individual biomarker cut-off.

Second, hypoxia-specific detection of biomarker release can be achieved through detection of 2-nitroimidazole modifications on proteins. Future studies should evaluate this concept in cellular systems of hypoxia and ultimately in patients presenting to a cardiac emergency unit. Moreover, this concept can be extended to other clinical situations in which detection of hypoxia-induced biomarker release is beneficial.

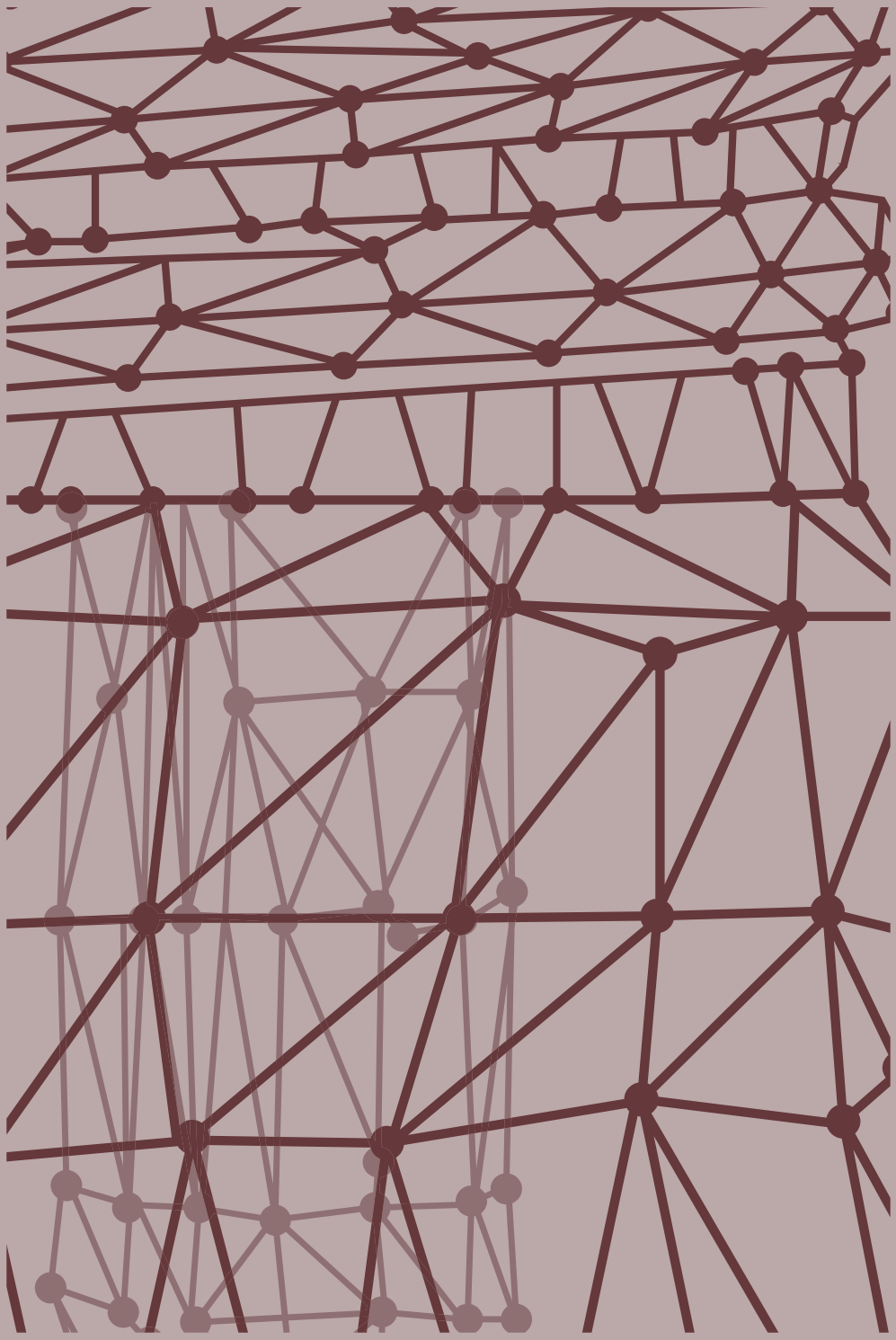
Third, combining multiple biomarkers into a clinical interpretable output through machine learning models is feasible and potentially proven to be clinically useful. This thesis demonstrated two examples with risk stratification in the emergency department and future glucose value prediction in individuals with diabetes. The scientific literature demonstrates similar findings, where retrospective development of machine learning models for almost any given medical prediction task is attainable. Hence, these initial findings now warrant further investigation in prospective, real-world implementation studies to evaluate the true clinical benefit of such technology.

Fourth, another interesting approach to optimize current use of biomarkers might be "biomarker re-purposing"[29]. This approach of applying known medical biomarkers for other purposes has been carried in pharmaceutical sciences, where recent large-scale, data-driven "drug repurposing" projects have been demonstrated[30].

References

1. Fuzery, A. K., Levin, J., Chan, M. M. & Chan, D. W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. Clin Proteomics 10, 13, doi:10.1186/1559-0275-10-13 (2013).
2. Katrukha, I. A. Human cardiac troponin complex. Structure and functions. Biochemistry (Mosc) 78, 1447-1465, doi:10.1134/S0006297913130063 (2013).
3. Neumann, J. T. et al. Application of High-Sensitivity Troponin in Suspected Myocardial Infarction. N Engl J Med 380, 2529-2540, doi:10.1056/NEJMoa1803377 (2019).
4. Reichlin, T. et al. Early diagnosis of myocardial infarction with sensitive cardiac troponin assays. N Engl J Med 361, 858-867, doi:10.1056/NEJMoa0900428 (2009).
5. Pickering, J. W. et al. Early kinetic profiles of troponin I and T measured by high-sensitivity assays in patients with myocardial infarction. Clin Chim Acta 505, 15-25, doi:10.1016/j.cca.2020.02.009 (2020).
6. Sandoval, Y. et al. Sex-Specific Kinetics of High-Sensitivity Cardiac Troponin I and T following Symptom Onset and Early Presentation in Non-ST-Segment Elevation Myocardial Infarction. Clin Chem, doi:10.1093/clinchem/hvaa263 (2020).
7. Katrukha, I. A. & Katrukha, A. G. Myocardial Injury and the Release of Troponins I and T in the Blood of Patients. Clin Chem 67, 124-130, doi:10.1093/clinchem/hvaa281 (2021).
8. De Wolf, H. A. et al. How well do laboratories adhere to recommended guidelines for dyslipidaemia management in Europe? The CArdiac MARker Guideline Uptake in Europe (CAMARGUE) study. Clin Chim Acta 508, 267-272, doi:10.1016/j.cca.2020.05.038 (2020).
9. Hammerer-Lercher, A. et al. Update on current practice in laboratory medicine in respect of natriuretic peptide testing for heart failure diagnosis and management in Europe. The CARdiac MArker guideline Uptake in Europe (CARMAGUE) study. Clin Chim Acta 511, 59-66, doi:10.1016/j.cca.2020.09.030 (2020).
10. Kimenai, D. M. et al. Ten Years of High-Sensitivity Cardiac Troponin Testing: Impact on the Diagnosis of Myocardial Infarction. Clin Chem, doi:10.1093/clinchem/hvaa272 (2020).
11. Collinson, P. et al. How Well Do Laboratories Adhere to Recommended Guidelines for Cardiac Biomarkers Management in Europe? The CArdiac MARker Guideline Uptake in Europe (CAMARGUE) Study of the European Federation of Laboratory Medicine Task Group on Cardiac Markers. Clin Chem, doi:10.1093/clinchem/hvab066 (2021).
12. Collinson, P. et al. How Well Do Laboratories Adhere to Recommended Clinical Guidelines for the Management of Myocardial Infarction: The CARdiac MArker Guidelines Uptake in Europe Study (CARMAGUE). Clin Chem 62, 1264-1271, doi:10.1373/clinchem.2016.259515 (2016).
13. Lowry, M. T. H., Anand, A. & Mills, N. L. Implementing an early rule-out pathway for acute myocardial infarction in clinical practice. Heart 107, 1912-1919, doi:10.1136/heartjnl-2019-316242 (2021).
14. Marini, M. G., Cardillo, M. T., Caroli, A., Sonnino, C. & Biasucci, L. M. Increasing specificity of high-sensitivity troponin: new approaches and perspectives in the diagnosis of acute coronary syndromes. J Cardiol 62, 205-209, doi:10.1016/j.jjcc.2013.04.005 (2013).
15. Chen, Y., Tao, Y., Zhang, L., Xu, W. & Zhou, X. Diagnostic and prognostic value of biomarkers in acute myocardial infarction. Postgrad Med J 95, 210-216, doi:10.1136/postgradmedj-2019-136409 (2019).
16. Vylegzhanina, A. V. et al. Full-Size and Partially Truncated Cardiac Troponin Complexes in the Blood of Patients with Acute Myocardial Infarction. Clin Chem 65, 882-892, doi:10.1373/clinchem.2018.301127 (2019).
17. Streng, A. S. et al. Identification and Characterization of Cardiac Troponin T Fragments in Serum of Patients Suffering from Acute Myocardial Infarction. Clin Chem 63, 563-572, doi:10.1373/clinchem.2016.261511 (2017).
18. Mair, J. et al. How is cardiac troponin released from injured myocardium? Eur Heart J Acute Cardiovasc Care, 2048872617748553, doi:10.1177/2048872617748553 (2017).
19. Mingels, A. M. et al. Cardiac Troponin T: Smaller Molecules in Patients with End-Stage Renal Disease than after Onset of Acute Myocardial Infarction. Clin Chem 63, 683-690, doi:10.1373/clinchem.2016.261644 (2017).

20. Vroemen, W. H. M. et al. Cardiac Troponin T: Only Small Molecules in Recreational Runners After Marathon Completion. J Appl Lab Med 3, 909-911, doi:10.1373/jalm.2018.027144 (2019).
21. Benjamens, S., Dhunnoo, P. & Mesko, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 3, 118, doi:10.1038/s41746-020-00324-0 (2020).
22. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ 368, m689, doi:10.1136/bmj.m689 (2020).
23. Cobelli, C., Renard, E. & Kovatchev, B. Artificial pancreas: past, present, future. Diabetes 60, 2672-2682, doi:10.2337/db11-0654 (2011).
24. Rodbard, D. Continuous Glucose Monitoring: A Review of Successes, Challenges, and Opportunities. Diabetes Technol Ther 18 Suppl 2, S3-S13, doi:10.1089/dia.2015.0417 (2016).
25. Woldaregay, A. Z. et al. Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. Artif Intell Med 98, 109-134, doi:10.1016/j.artmed.2019.07.007 (2019).
26. Martinsson, J. et al. in 3rd International Workshop on Knowledge Discovery in Healthcare Data, KDH@IJCAI-ECAI 2018, 13 July 2018 64-68 (2018).
27. Faruqui, S. H. A. et al. Development of a Deep Learning Model for Dynamic Forecasting of Blood Glucose Level for Type 2 Diabetes Mellitus: Secondary Analysis of a Randomized Controlled Trial. JMIR Mhealth Uhealth 7, e14452, doi:10.2196/14452 (2019).
28. Marling, C. & Bunescu, R. C. in KHD@IJCAI.
29. Ohmichi, T. et al. Biomarker repurposing: Therapeutic drug monitoring of serum theophylline offers a potential diagnostic biomarker of Parkinson's disease. PLoS One 13, e0201260, doi:10.1371/journal.pone.0201260 (2018).
30. Rodriguez, S. et al. Machine learning identifies candidates for drug repurposing in Alzheimer's disease. Nat Commun 12, 1033, doi:10.1038/s41467-021-21330-0 (2021).



SUPPLEMENTALS

SUPPLEMENTALS

This chapter describes the valorization of the research performed in this thesis. Valorization is hereby defined as the social and economic relevance of research, the target populations of the findings, the related concrete products, services, processes and activities, their innovativeness and future directions[1].

Research field and main objectives

This thesis built on alternative strategies to address unmet diagnostic challenges, by optimizing current biomarkers, rather than developing new ones. This optimization was examined for several diagnostic challenges, each with its corresponding clinical target population. The first clinical population suffered from cardiovascular disease. Despite the decreasing prevalence of cardiovascular disease in the Western world, myocardial infarction (MI) and heart failure (HF) remain an important burden of morbidity and mortality worldwide[2,3]. In the Netherlands, consisting of approximately 17 million individuals, each year 75.000 and 240.000 people are diagnosed with myocardial infarction and heart failure, respectively[4,5]. Accurate and rapid discrimination of patients with or without these pathologies facilitates appropriate treatment and prioritization of resources[6-9]. The second population that was studied involved patients presenting to the emergency department (ED). The number of patients referred to ED are increasing worldwide[10,11]. Prolonged waiting times and associated crowding in the ED increase mortality up to 10%[12], and rapid triage is therefore a core task in emergency medicine. An effective means to identify patients at high- and low-risk shortly after admission could help decision-making regarding patient prioritization, treatment, level of observation, and post-discharge follow-up. The third population included individuals with diabetes mellitus (DM), a metabolic disease that is characterized by elevated blood glucose values. At present, approximately 463 million individuals are affected by diabetes worldwide; a figure that is expected to rise to 700 million by 2045[13]. This is problematic because diabetes lowers life expectancy[14,15] and strongly increases the chances of several diseases, including cardiovascular disease[15], eye, kidney, and nerve disease[13], and dementia[16].

Relevance and utilization of the key findings

The current thesis describes several research findings that are relevant and could be utilized to address unmet diagnostic challenges. Three key research findings are highlighted in this section.

Rapid discrimination of hypoxia-induced troponin release through 2-nitroimidazole detection.
The clinical performance of the current immunoassay for cardiac troponin exhibits high sensitivity and negative predictive value (>99%) but suboptimal specificity and positive predictive value (±75%) for the diagnosis of MI. The suboptimal specificity and PPV are partly caused by non-hypoxic pathologies that can also increase cardiac troponins. Hence, an unmet clinical need became a cardiac biomarker assay that could discriminate

between hypoxia-induced cardiomyocyte damage (AMI) and cardiomyocyte damage other than AMI. An innovative tool was introduced in Chapter 6 allowing the selective detection of cardiac troponin released under conditions of hypoxia, but leaving biomarker release due to non-hypoxic injury triggers undetected. In particular, a family of compounds called 2-nitroimidazole were demonstrated to selectively modify proteins under conditions of hypoxia, leading to the development of a novel mass spectrometry assay that detects 2-nitroimidazole modifications on proteins (such as troponin). These research findings should further be validated in cellular systems of hypoxia and ultimately in patients requiring rapid discrimination between hypoxia-induced cardiomyocyte damage (AMI) and cardiomyocyte damage other than AMI. Expedited rule-in or rule-out of these patients will improve patient outcome, reduce anxiety for patients, and improve work flows for cardiologists in cardiac emergency units. Importantly, the proposed concept is not limited to the diagnosis of MI, but should also be of interest to clinicians facing similar clinical challenges where rapid discrimination between hypoxic versus non-hypoxic cellular necrosis is clinically relevant, such as in the detection of renal ischemia using neutrophil gelatinase-associated lipocalin (NGAL).

Rapid and accurate risk stratification at the emergency department through combination of biomarkers. Numerous clinical risk scores and triage systems for stratification of patients in the ED have been developed. Unfortunately, these systems often generalize poorly and lack precision, impeding their clinical use. In this thesis, demographical and laboratory data was employed to develop a novel clinical decision tool with machine learning outperforming established clinical risk scores and internal medicine physicians (Chapters 7 to 9). Specifically, the novel decision support tool was able to accurately predict a patient's 31-day mortality risk. These research findings provide a solid basis for prospective, follow-up studies evaluating the performance and clinical benefit of these models in a real-world, clinical emergency department setting. Ultimately this clinical decision support tool might provide beneficial in reducing length-of-stay or more optimal management of these patients. Moreover, these studies extensively reflected on the methodology and techniques used for mortality prediction, with the aim of providing fundamental, technical insights that can help improve future research related to machine-learning based risk prediction in medicine.

Next-generation closed-loop insulin dosing systems through anticipation on future glucose levels. One of the most promising developments to better regulate glucose levels in individuals with diabetes who require insulin treatment is a closed-loop insulin delivery. Such a system integrates continuous glucose monitoring (CGM), insulin (with or without glucagon) infusion, and a control algorithm to continuously regulate blood glucose levels. These devices may be improved by addressing certain inherent shortcomings related to CGM, such as sensor delay and brief periods of sensor malfunction24, 30, 31. Additionally,

anticipation on future glucose values may improve these devices even further. In this thesis, historical glucose levels and physical activity data was combined to predict future glucose values in order to improve the CGM part of closed-loop insulin delivery systems. Our model was able to accurately and safely predict glucose values at 15- and 60-minute intervals, which could be useful in case of short periods of CGM malfunction and further improvement of these systems. Future studies should explore the possibility of further optimizing these prediction models in cooperation with companies that are specialized in diabetes care. Moreover, our findings contribute to knowledge related to different preprocessing steps, algorithm choices and utility of transfer learning techniques in the field of glucose prediction.

Knowledge dissemination to target groups

The impact of this thesis can be considered to be primarily scientific but is also of importance to clinicians and the (diagnostic) laboratory industry. The scientific disciplines that can benefit from the findings described in this thesis are diverse; for example, the novel methodology for detection of hypoxia-induced biomarker release was proposed, is mostly interesting for disciplines such as (molecular) cell biology, biomarker research and (clinical) cardiology. On the other hand, the newly developed clinical decision support tools are not only relevant for scientists and clinicians active in that particular field (e.g., emergency medicine in case of risk stratification or endocrinology in case of glucose prediction) but also for fundamental data and computer scientists that aim to evolve these tools technically. In general, publication of the results in scientific journals -both preprint or peer-reviewed- was one of the main ways to inform this community. In addition, the results of this thesis have been presented at several national and international conferences.

The findings presented in this thesis should also be of particular interest to the diagnostic and laboratory industry. The thesis presents several opportunities to translate research findings into commercially available diagnostic tools; for example, the development of an hypoxia-specific assay for cardiac troponin or the integration of the risk stratification tools into laboratory information systems.

References

1. Universiteiten, V. v. Raamwerk valorisatie-indicatoren., 2013).
2. Townsend, N. et al. Cardiovascular disease in Europe: epidemiological update 2016. Eur Heart J 37, 3232-3245, doi:10.1093/eurheartj/ehw334 (2016).
3. Mortality, G. B. D. & Causes of Death, C. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 388, 1459-1544, doi:10.1016/S0140-6736(16)31012-1 (2016).
4. Volksgezondheidenzorg.info.Volksgezondheidenzorg.info, <Volksgezondheidenzorg.info> (2016).
5. OECD Better Life Index, 2016).
6. Neumann, J. T. et al. Application of High-Sensitivity Troponin in Suspected Myocardial Infarction. N Engl J Med 380, 2529-2540, doi:10.1056/NEJMoa1803377 (2019).
7. Westwood, M. E. et al. Optimizing the Use of High-Sensitivity Troponin Assays for the Early Rule-out of Myocardial Infarction in Patients Presenting with Chest Pain: A Systematic Review. Clin Chem 67, 237-244, doi:10.1093/clinchem/hvaa280 (2021).
8. Long, B., Koyfman, A. & Gottlieb, M. Diagnosis of Acute Heart Failure in the Emergency Department: An Evidence-Based Review. West J Emerg Med 20, 875-884, doi:10.5811/westjem.2019.9.43732 (2019).
9. Pourafkari, L., Tajlil, A. & Nader, N. D. Biomarkers in diagnosing and treatment of acute heart failure. Biomark Med 13, 1235-1249, doi:10.2217/bmm-2019-0134 (2019).
10. Hooker, E. A., Mallow, P. J. & Oglesby, M. M. Characteristics and Trends of Emergency Department Visits in the United States (2010-2014). J Emerg Med 56, 344-351, doi:10.1016/j.jemermed.2018.12.025 (2019).
11. Wansink, L. et al. Trend analysis of emergency department malpractice claims in the Netherlands: a retrospective cohort analysis. Eur J Emerg Med 26, 350-355, doi:10.1097/MEJ.0000000000000572 (2019).
12. Guttmann, A., Schull, M. J., Vermeulen, M. J. & Stukel, T. A. Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada. BMJ 342, d2983, doi:10.1136/bmj.d2983 (2011).
13. Low Wang, C. C., Hess, C. N., Hiatt, W. R. & Goldfine, A. B. Clinical Update: Cardiovascular Disease in Diabetes Mellitus: Atherosclerotic Cardiovascular Disease and Heart Failure in Type 2 Diabetes Mellitus - Mechanisms, Management, and Clinical Considerations. Circulation 133, 2459-2502, doi:10.1161/CIRCULATIONAHA.116.022194 (2016).
14. American Diabetes, A. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2020. Diabetes Care 43, S14-S31, doi:10.2337/dc20-S002 (2020).
15. Preis, S. R. et al. Trends in all-cause and cardiovascular disease mortality among women and men with and without diabetes mellitus in the Framingham Heart Study, 1950 to 2005. Circulation 119, 1728-1735, doi:10.1161/CIRCULATIONAHA.108.829176 (2009).
16. Stehouwer, C. D. A. Microvascular Dysfunction and Hyperglycemia: A Vicious Cycle With Widespread Consequences. Diabetes 67, 1729-1741, doi:10.2337/dbi17-0044 (2018).

SUPPLEMENTALS

Summary (English)

The majority of day-to-day clinical decision making is guided by the laboratory measurement of biomarkers. Many biomarkers to date provide benefit in this day-to-day decision making, but there is still considerable need for more optimal biomarkers. Despite monumental technological advances in biomarker discovery, it is estimated that on average less than one novel biomarker is approved by the regulations on a yearly basis. This thesis examines characteristics of current acute (cardiac) care biomarkers, and subsequently demonstrates various approaches to optimize current biomarkers rather than deploying new ones as an attempt to improve day-to-day clinical decision making.

Chapter 1 provides an introduction to the process of biomarker discovery, implementation and evaluation. It furthermore reviews the large discrepancy between the number of newly discovered biomarkers versus the number of biomarkers actually translating into clinical practice. Based on this observation, alternative approaches –relying on optimizing current biomarkers, rather than deploying new ones- are reviewed to improve the diagnostic trajectory.

Chapter 2 reviews the analytical characteristics of the biochemical cornerstone in the diagnosis of myocardial infarction (MI); cardiac troponin. Specifically, this chapter provides clinical laboratory practice recommendations for high-sensitivity cardiac troponin (cTn) testing. In addition to the recommendations as described by others, several key aspects are discussed such as troponin cut-off values, differences between the United States (US) and Europe in regard to assay characteristics and importance of blood matrix selection in relation to troponin molecular degradation. This chapter emphasizes the importance of guidance, education and support of clinicians by laboratory specialists when implementing a novel biomarker. Although cardiac troponin is often referred to as a single biomarker, it is a protein complex existing of three proteins of which two are of clinical relevance: cardiac troponin T (cTnT) and cardiac troponin I (cTnI). The current evidence in relation to clinical performance suggests that cTnT and cTnI can be considered equally good.

Chapter 3 examines the kinetics for troponin T and troponin I in patients with a non-ST-segment elevated MI (NSTEMI). Previous research demonstrated distinct kinetics between troponin I (monophasic) and troponin T (biphasic) in patients with STEMI. However, this is an MI population that more relies on electrocardiography rather than biochemical troponin measurements. The troponin kinetics in NSTEMI confirmed the monophasic troponin I and biphasic troponin T curve. It moreover appeared that baseline and peak troponin I concentrations were substantially higher compared to troponin T. Despite multiple hypotheses, the difference in kinetics between the two cardiac troponins remains to be explored.

Chapter 4 further evaluates the analytical characteristics of the troponin biomarker by assessing the real-world impact of biotin interference in an acute cardiac care setting. A recent safety warning by the US Food and Drug Administration (FDA) suggested that troponin concentrations can be falsely low due to endogenous biotin through dietary supplement intake. In a large acute cardiac care population it was demonstrated that biotin interference had a negligible effect on hs-cTnT concentrations by successfully depleting endogenous biotin. Thus, the absolute effect of biotin interference on the analytical characteristics is lower than other common interference sources such as hemolysis and blood sample misidentification.

Chapter 5 studies the impact of a potential circadian rhythm of various natriuretic peptides (NP) on the clinical performance for the diagnosis of AHF. Natriuretic peptides are the biochemical gold standard for the diagnosis of. Current guidelines assume random variation of natriuretic peptides and therefore recommend interpreting concentrations irrespective of the time of emergency department presentation. This chapter provides substantial evidence for circadian rhythms of some of these natriuretic peptides, with lower concentrations during evening/nights and higher concentrations during daytime. Despite statistically impacting the clinical performance for AHF depending on presentation time, the diagnostic accuracy of these natriuretic peptides for AHF remained very high at all times.
To overcome the challenges associated with deploying a new biomarker in routine clinical care, this thesis pursues an alternative approach to improve the day-to-day clinical decision making by optimizing current biomarkers, rather than deploying new ones. Such approach may benefit from lower costs, existing infrastructure, established knowledge in relation to the biomarker biology and detection, and more convenient implementation as the biomarker is likely part of clinical guidelines and thus known by clinicians and laboratory specialists.

Chapter 6 proposes a novel concept to improve the diagnostic specificity of the cardiac biomarker troponin. Despite troponin's excellent sensitivity and cardiac tissue specificity, the diagnosis of MI often remains challenging. This complexity stems from the inability of troponin to discriminate between troponin elevations due to hypoxic cardiac injury and troponin elevations from other causes of cardiac cell death. Combining tissue specificity with hypoxia in a single biomarker is therefore an unmet clinical need, but as a native molecule, such marker may not even exist. In Chapter 6, we unraveled the precise configuration of the hypoxia-specific 2-nitroimidazole modification on proteins. The results in this chapter elucidate the binding mechanism of 2-nitroimidazoles to proteins, and provide the molecular basis for the discrimination of hypoxic versus non-hypoxic cell death in protein biomarker assays. Based on these findings, a proof-of-concept mass spectrometry assay was develop to detect hypoxia-dependent troponin release.

The last **four chapters** employ machine learning to combine multiple biomarker results into a clinically relevant output as a means to develop a clinical decision support system, and thereby improve day-to-day clinical decision making.

Chapter 7 demonstrates the development of a clinical decision support tool based upon a profile of biomarker levels in sepsis patients presenting to the emergency department. Laboratory and clinical biomarker data was combined through machine learning to predict 31-day mortality in these patients. The clinical decision support tool outperformed established clinical risk scores and internal medicine specialists in a direct comparison.

The application of this clinical decision support tool was extended in **Chapter 8,** by developing and evaluating its performance in general emergency department populations across four large Dutch hospitals. Moreover, the tool was designed to provide information about importance of biomarker levels in a single prediction, thereby partly mitigating the issue of 'black-box' machine learning. The clinical decision support tool demonstrated a consistent high performance for predicting mortality at 31 days (AUC 0.87 and higher), highly outperforming any clinical risk score to date. Follow-up studies should now assess whether implementation of this clinical decision support tool can improve clinically relevant endpoints. As a result of the recent COVID-19 pandemic, the clinical decision support tool was also validated in a small subgroup of patients with COVID-19 in **Chapter 9**. The discriminatory performance of the tool was lower than previously reported in a general population, but still among the highest performing clinical risk scores for predicting 31-day mortality in patients with COVID-19.

In **Chapter 10**, historical glucose and physical activity biomarker data were combined in healthy individuals and individuals with (pre-)diabetes to predict their future glucose values on a 15- and 60-minute interval. It was demonstrated that predictions were not only accurate in individuals with type 1 and 2 diabetes, but also clinically safe as demonstrated by surveillance error grids. This clinical decision support tool could potentially aid in overcoming inherent shortcomings of the current closed-loop insulin dosing systems.

Lastly, **Chapter 11,** contains a general discussion of the work presented in this thesis and provides directions for future research. The current thesis reviewed the analytical and clinical performance of current acute (cardiac) care biomarkers. Moreover, this thesis explored various approaches to optimize current biomarkers rather than developing new ones. Future research should further examine these approaches and study their beneficial value in day-day-day clinical decision making.

SUPPLEMENTALS

Samenvatting (Nederlands)

Het merendeel van de dagelijkse klinische besluitvorming wordt gestuurd door de laboratoriummeting van biomarkers. Veel biomarkers zijn tot op heden nuttig voor deze dagelijkse besluitvorming, maar er is nog steeds een aanzienlijke behoefte aan betere, accuratere en goedkopere biomarkers. Ondanks monumentale technologische vooruitgangen in de ontdekking van biomarkers, wordt naar schatting jaarlijks gemiddeld slechts minder dan één nieuwe biomarker goedgekeurd door de regelgeving. Dit proefschrift onderzoekt kenmerken van de huidige biomarkers voor acute (cardiale) zorg, en toont vervolgens verschillende benaderingen om de huidige biomarkers te optimaliseren in plaats van nieuwe in te zetten in een poging om de dagelijkse klinische besluitvorming te verbeteren.

Hoofdstuk 1 geeft een inleiding tot het proces van ontdekking, implementatie en evaluatie van biomarkers. Verder wordt in gegaan op de grote discrepantie tussen het aantal nieuw ontdekte biomarkers en het aantal biomarkers dat daadwerkelijk in de klinische praktijk wordt toegepast. Op basis van deze observatie worden alternatieve benaderingen - die gebaseerd zijn op het optimaliseren van bestaande biomarkers in plaats van op het inzetten van nieuwe - besproken om het diagnostische traject van de huidige biomarkers te verbeteren.

Hoofdstuk 2 beschrijft de analytische kenmerken van de biochemische hoeksteen in de diagnose van myocardinfarct (MI); cardiaal troponine. Dit hoofdstuk bevat aanbevelingen voor de klinische laboratoriumpraktijk met betrekking tot de hoog-gevoelige ('high-sensitive') cardiaal troponine test (hs-cTn). Naast de aanbevelingen zoals door anderen beschreven, worden verschillende belangrijke aspecten besproken zoals de afkapwaarden voor troponine, verschillen tussen de Verenigde Staten (VS) en Europa met betrekking tot assay-kenmerken en het belang van de selectie van de bloedmatrix in relatie tot de moleculaire afbraak van troponine. Dit hoofdstuk benadrukt het belang van begeleiding, opleiding en ondersteuning van clinici door laboratoriumspecialisten bij de toepassing van een nieuwe biomarker. Hoewel cardiaal troponine vaak wordt aangeduid als één biomarker, is het een eiwitcomplex dat bestaat uit drie eiwitten waarvan er twee van klinisch belang zijn: cardiaal troponine T (cTnT) en cardiaal troponine I (cTnI). Het huidige bewijs met betrekking tot klinische prestaties suggereert dat cTnT en cTnI als even goed kunnen worden beschouwd.

Hoofdstuk 3 onderzoekt de kinetiek voor troponine T en troponine I bij patiënten met een niet-ST-segment verhoogd MI (NSTEMI). Eerder onderzoek toonde verschillende kinetieken aan voor troponine I (monofasisch) en troponine T (bifasisch) bij patiënten met STEMI. Dit is echter een groep patiënten die meer steunt op elektrocardiografie voor diagnostiek dan op troponine metingen. De kinetiek van troponine in NSTEMI bevestigde

de monofasische troponine I en bifasische troponine T curve. Bovendien bleek dat de basislijn- en piekconcentraties van troponine I aanzienlijk hoger waren in vergelijking met troponine T. Ondanks meerdere hypothesen moet het verschil in kinetiek tussen de twee cardiale troponines nog worden onderzocht.

Hoofdstuk 4 gaat verder in op de analytische kenmerken van de troponine-biomarker door de reële impact van biotine-interferentie binnen patiënten afkomstig van de eerste hart hulp (EHH) te beoordelen. Een recente veiligheidswaarschuwing van de Amerikaanse Food and Drug Administration (FDA) suggereerde dat troponine concentraties ten onrechte laag kunnen zijn als gevolg van endogene biotine door inname van voedingssupplementen. In een grote populatie patiënten afkomstig van de eerste hart hulp werd aangetoond dat biotine-interferentie een verwaarloosbaar effect had op hs-cTnT-concentraties door met succes endogene biotine te depleteren. Het absolute effect van biotine-interferentie op de analytische kenmerken is dus kleiner dan andere veel voorkomende interferentiebronnen zoals hemolyse en verkeerde identificatie van bloedmonsters.

Hoofdstuk 5 onderzoekt de invloed van een mogelijk circadiaans ritme van verschillende natriuretische peptiden (NP) op de klinische prestaties voor de diagnose van acuut hart falen (AHF). Natriuretische peptiden zijn de biochemische gouden standaard voor de diagnose van AHF. De huidige richtlijnen gaan uit van willekeurige variatie van natriuretische peptiden en bevelen daarom aan de concentraties te interpreteren ongeacht het tijdstip van presentatie op de spoedeisende hulp. Dit hoofdstuk levert substantieel bewijs voor circadiane ritmes van sommige van deze natriuretische peptiden, met lagere concentraties ‘s avonds/nachten en hogere concentraties overdag. Ondanks een statistisch effect op de klinische prestaties voor AHF afhankelijk van het tijdstip van presentatie, bleef de diagnostische accuratesse van deze natriuretische peptiden voor AHF op elk moment zeer hoog.

Om de uitdagingen te overwinnen die gepaard gaan met het inzetten van een nieuwe biomarker in de routine klinische zorg, wordt in dit proefschrift een alternatieve aanpak gevolgd om de dagelijkse klinische besluitvorming te verbeteren door het optimaliseren van bestaande biomarkers, in plaats van het inzetten van nieuwe. Een dergelijke aanpak kan profiteren van lagere kosten, bestaande infrastructuur, gevestigde kennis met betrekking tot de biologie en detectie van de biomarker, en een gemakkelijkere implementatie omdat de biomarker waarschijnlijk deel uitmaakt van klinische richtlijnen en dus bekend is bij clinici en laboratoriumspecialisten.

In **hoofdstuk 6** wordt een nieuw concept voorgesteld om de diagnostische specificiteit van de cardiale biomarker troponine te verbeteren. Ondanks de uitstekende gevoeligheid en specificiteit van troponine blijft de diagnose van MI vaak een uitdaging. Deze complexiteit komt voort uit het onvermogen van troponine om onderscheid te maken tussen troponine verhogingen als gevolg van hypoxisch hartletsel en troponine verhogingen als gevolg van andere oorzaken van een hartinfarct celdood. Het combineren van weefselspecificiteit met hypoxie in één enkele biomarker is daarom een onvervulde klinische behoefte, maar als een natief molecule bestaat zo'n marker waarschijnlijk niet. In **hoofdstuk 6** ontrafelden we de exacte configuratie van de hypoxie-specifieke 2-nitroimidazol modificatie op eiwitten. De resultaten in dit hoofdstuk verhelderen het bindingsmechanisme van 2-nitroimidazolen aan eiwitten, en verschaffen de moleculaire basis voor het onderscheid tussen hypoxische en niet-hypoxische celdood in eiwit biomarker assays. Op basis van deze bevindingen werd een proof-of-concept massaspectrometrietest ontwikkeld om hypoxie-afhankelijke afgifte van troponine te detecteren.

In de laatste **vier hoofdstukken** wordt machine learning (ML) toegepast om meerdere biomarker resultaten te combineren tot een klinisch relevante output als middel om een klinisch beslissingsondersteunend systeem te ontwikkelen, en daarmee de dagelijkse klinische besluitvorming te verbeteren.

Hoofdstuk 7 demonstreert de ontwikkeling van een hulpmiddel voor klinische besluitvorming op basis van een profiel van biomarkers bij patiënten met sepsis die zich presenteerden op de spoedeisende hulp. De laboratorium- en klinische biomarker-gegevens werden door middel van machine learning gecombineerd om de 31-daagse mortaliteit bij deze patiënten te voorspellen. De klinische beslissingsondersteunende tool presteerde beter dan gevestigde klinische risicoscores en interne geneeskunde specialisten in een directe vergelijking.

De toepassing van deze klinische beslissingsondersteunende tool werd uitgebreid in **hoofdstuk 8**, door de ontwikkeling en evaluatie van de prestaties in populaties van vier spoedeisende hulpen afkomstig uit vier grote Nederlandse ziekenhuizen. Bovendien werd het instrument ontworpen om informatie te verstrekken over het belang van biomarker niveaus in een enkele voorspelling, waardoor het probleem van 'black-box' machine learning gedeeltelijk werd ondervangen. De klinische beslissingsondersteunende tool liet een consistent hoge prestatie zien voor het voorspellen van sterfte na 31 dagen (AUC 0,87 en hoger), waarmee het veel beter presteerde dan welke klinische risicoscore dan ook. In vervolgstudies moet nu worden nagegaan of de toepassing van dit hulpmiddel voor klinische besluitvorming de klinisch relevante eindpunten kan verbeteren.

Naar aanleiding van de recente COVID-19 pandemie werd het hulpmiddel voor klinische besluitvorming ook gevalideerd in een kleine subgroep van patiënten met COVID-19 in **hoofdstuk 9**. De discriminerende prestaties van het instrument waren lager dan eerder gerapporteerd in een algemene populatie, maar behoorden nog steeds tot de best presterende klinische risicoscores voor het voorspellen van 31-daagse mortaliteit bij patiënten met COVID-19.

In **hoofdstuk 10** werden historische glucose en fysieke activiteit biomarker gegevens gecombineerd bij gezonde personen en personen met (pre-)diabetes om hun toekomstige glucosewaarden te voorspellen met een interval van 15 en 60 minuten. Aangetoond werd dat de voorspellingen niet alleen accuraat waren bij personen met diabetes type 1 en 2, maar ook klinisch veilig, zoals blijkt uit verschillende veiligheidsanalyses. Dit hulpmiddel ter ondersteuning van klinische beslissingen kan mogelijk helpen bij het ondervangen van inherente tekortkomingen van de huidige gesloten-lus-systemen voor insulinedosering.

Hoofdstuk 11 tenslotte bevat een algemene discussie van het in dit proefschrift gepresenteerde werk en geeft aanwijzingen voor toekomstig onderzoek. In dit proefschrift werden de analytische en klinische prestaties van de huidige biomarkers voor acute (hart)zorg geëvalueerd. Bovendien werden in dit proefschrift verschillende benaderingen onderzocht om de huidige biomarkers te optimaliseren in plaats van nieuwe te ontwikkelen. Toekomstig onderzoek moet deze benaderingen verder onderzoeken en hun gunstige waarde voor de dagelijkse klinische besluitvorming bestuderen.

SUPPLEMENTALS

List of abbreviations

AHF	acute heart failure
AUC	area under the curve
BNP	B-type natriuretic peptide
CDS	clinical decision support
CGM	continuous glucose monitoring
CI	confidence intervals
CKD-EPI	chronic kidney disease epidemiology collaboration formula
cTn	cardiac troponin
cTnI	cardiac troponin I
cTnT	cardiac troponin T
CV	coefficient of variation
ED	emergency department
FBS	fetal bovine serum
hs-cTnI	high-sensitivity cardiac troponin I
hs-cTnT	high-sensitivity cardiac troponin T
ICU	intensive care unit
IQR	interquartile range
kDa	kilo Dalton
LC-MS	liquid chromatography - mass spectrometry
LightGBM	light gradient boosting system
LSTM	long-short term memory
LoB	limit of blank
LoD	limit of detection
LoQ	limit of quantification
mAb	monoclonal antibody
ML	machine learning
MI	myocardial infarction
MS	mass spectrometry
NGM	normal glucose metabolism
NI	nitroimidazole
NP	natriuretic peptide
NPV	negative predictive value
NSTEMI	non-ST-elevation myocardial infarction
NT-proBNP	N-terminal pro-BNP
PreD	prediabetes
PPV	positive predictive value
QC	quality control
SD	standard deviation

SHAP	shapley additive explanations
STEMI	ST-elevation myocardial infarction
RNN	recurrent neural network
RMSE	root-mean-square error
T2D	type 2 diabetes
TnC	troponin C
URL	upper reference limit
WB	Western blot
XGBoost	eXtreme Gradient Boosting system

SUPPLEMENTALS

Thesis publications

van Doorn WPTM, Kadambar VK, van de Laak J, Kocken JMM, Streng AS, Deckers N, Schurgers LJ, Rouschop KMA, Bouwman FG, Melman A, Darie CC, Bekers O, Koch CJ, Meex SJR. Towards a hypoxia specific troponin assay: characterization of nitroimidazole-protein adducts. *Submitted*.

van Dam PMEL, **van Doorn WPTM**, Meex SJR, Stassen PM. Machine learning for risk stratification in patients with COVID-19 in the emergency department. *In preparation*.

van Doorn WPTM, Helmich F, Dam PMEL van, Jacobs LHJ, Stassen PM, Bekers O, Meex SJR. Explainable Machine Learning models for Rapid Risk Stratification in the Emergency Department: A multi-center study. medRxiv 2021; : 2020.11.25.20238386.

Breidthardt T*, **van Doorn WPTM***, van der Linden N, Diebold M, Wussler D, Danier I, Zimmermann T, Shrestha S, Kozhuharov N, Belkin M, Porta C, Strebel I, Michou E, Gualandro DM, Nowak A, Meex SJR, Mueller C. Diurnal Variations in Natriuretic Peptide Levels: Clinical Implications for the Diagnosis of Acute Heart Failure. Circ Heart Fail 2022; 15: e009165.

van Doorn WPTM*, Foreman YD*, Schaper NC, Savelberg HHCM, Koster A, van der Kallen CJH, Schram MT, Henry RMA, Dagnelie PC, de Galan BE, Bekers O, Stehouwer CDA, Meex SJR, Brouwers MCGJ. Machine learning-based glucose prediction with use of continuous glucose and physical activity monitoring data: The Maastricht Study. PLoS One 2021; 16: e0253125.

van Doorn WPTM, Stassen PM, Borggreve HF, Schalkwijk MJ, Stoffers J, Bekers O, Meex SJR. A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. PLoS ONE 2021; 16: e0245157.

Vroemen WHM*, **van Doorn WPTM***, Kimenai DM, Wodzig WKWH, de Boer D, Bekers O, Meex SJR. Biotin interference in high-sensitivity cardiac troponin T testing: a real-world evaluation in acute cardiac care. Cardiovascular Research 2019; 115: 1950–1951.

van Doorn WPTM*, Vroemen WHM*, Smulders MW, van Suijlen JD, van Cauteren YJM, Bekkers SCAM, Bekers O, Meex SJR. High-Sensitivity Cardiac Troponin I and T Kinetics after Non-ST-Segment Elevation Myocardial Infarction. The Journal of Applied Laboratory Medicine 2020; 5: 239–241.

van Doorn WPTM*, Vroemen WHM, de Boer D, Mingels AMA, Bekers O, Wodzig WKWH, Meex SJR. Clinical laboratory practice recommendations for high-sensitivity cardiac troponin testing. J Lab Precis Med 2018; 3: 30–30.

Other publications

Denessen EJ, Heuts S, Daemen JH, **van Doorn WPTM**, Vroemen WHM, Sels JW, Segers P, Van t HOF AW, Maessen JG, Bekers O, van der Horst IC, Mingels AM. High-Sensitivity Cardiac Troponin I and T Kinetics Differ following Coronary Bypass Surgery: A Systematic Review and Meta-Analysis. Clinical Chemistry 2022; : hvac152.

Mimpen M, Damoiseaux J, **van Doorn WPTM**, Rolf L, Muris A-H, Hupperts R, van Luijn MM, Gerlach O, Smolders J. Proportions of circulating transitional B cells associate with MRI activity in interferon beta-treated multiple sclerosis patients. J Neuroimmunol 2021; 358: 577664.

Foreman YD, **van Doorn WPTM**, Schaper NC, van Greevenbroek MMJ, van der Kallen CJH, Henry RMA, Koster A, Eussen SJPM, Wesselius I, Reesink KD, Schram MT, Dagnelie PC, Kroon AA, Brouwers MCGJ, Stehouwer CDA. Greater daily glucose variability and lower time in range assessed with continuous glucose monitoring are associated with greater aortic stiffness: The Maastricht Study. Diabetologia 2021; 64: 1880–1892.

de Boer D, Streng AS, **van Doorn WPTM**, Vroemen WHM, Bekers O, Wodzig WKWH, Mingels AMA. Cardiac Troponin T: The Impact of Posttranslational Modifications on Analytical Immunoreactivity in Blood up to the Excretion in Urine. Adv Exp Med Biol 2021; 1306: 41–59.

Damoiseaux M, **van Doorn WPTM**, van Lochem E, Damoiseaux J. Testing for IgA anti-tissue transglutaminase in routine clinical practice: Requesting behaviour in relation to prevalence of positive results. J Transl Autoimmun 2020; 3: 100045.

Hendrix M, Bons J, van Haren A, van Kuijk S, **van Doorn WPTM**, Kimenai DM, Bekers O, Spaanderman M, Al-Nasiry S. Role of sFlt-1 and PlGF in the screening of small-for-gestational age neonates during pregnancy: A systematic review. Ann Clin Biochem 2020; 57: 44–58.

van der Linden N, Hilderink JM, Cornelis T, Kimenai DM, Klinkenberg LJJ, **van Doorn WPTM**, Litjens EJR, van Suijlen JD, van Loon LJC, Bekers O, Kooman JP, Meex SJR. Twenty-Four-Hour Biological Variation Profiles of Cardiac Troponin I in Individuals with or without Chronic Kidney Disease. Clin Chem 2017; 63: 1655–1656.

Streng AS, de Boer D, **van Doorn WPTM**, Bouwman FG, Mariman ECM, Bekers O, van Dieijen-Visser MP, Wodzig WKWH. Identification and Characterization of Cardiac Troponin T Fragments in Serum of Patients Suffering from Acute Myocardial Infarction. Clin Chem 2017; 63: 563–572.

Streng AS, de Boer D, **van Doorn WPTM**, Kocken JMM, Bekers O, Wodzig WKWH. Cardiac troponin T degradation in serum is catalysed by human thrombin. Biochem Biophys Res Commun 2016; 481: 165–168.

SUPPLEMENTALS

About the author

William van Doorn was born on June 11, 1994 in Heesch, The Netherlands. After graduating from secondary school (HAVO) at Mondriaan College (Oss, the Netherlands) in 2011, he studied Chemistry at Avans University of Applied Sciences (Breda, the Netherlands). After obtaining his Bachelor of Science in 2015, William continued with a master programme in Biomedical Sciences at Maastricht University (Maastricht, the Netherlands) with a focus on cardiovascular medicine. During his master's programme, William worked as a parttime research assistant in the Central Diagnostic Laboratory under supervision of Dr. Meex and Prof. Dr. Otto Bekers. After finishing his final master internship at the same department studying the binding of nitroimidazoles in cell systems during hypoxia, William obtained his Master of Science in 2017. After his graduation, he started as a PhD-student under supervision of prof. dr. Otto Bekers and dr. Steven Meex at the department of Clinical Chemistry within the Central Diagnostic Laboratory (Maastricht, the Netherlands). The results of the PhD trajectory are presented in this thesis. William started as a resident in clinical chemistry at the Maastricht UMC+ in 2021.

SUPPLEMENTALS

Dankwoord

Prof. dr. Bekers, beste **Otto**, hartelijk dank dat ik onder jouw leiding binnen het Centraal Diagnostisch Laboratorium (CDL) mocht promoveren. Jouw hulp en advies tijdens cardio meetings, reviews van wetenschappelijk artikelen en de mogelijkheid tot het volgen van cursussen en congressen waren zeer waardevol. Daarnaast heb ik enorm mogen genieten van je nuchtere en relativerende blik op het leven.

Dr. Meex, beste **Steven**, ik leerde je voor het eerst kennen toen ik een student-assistentschap onder je mocht uitvoeren tijdens mijn master. Ik heb bewondering voor je, zowel binnen als buiten het werk. Naast dat ik wetenschappelijk heel veel van je heb mogen leren, heb ik daarbuiten me waarschijnlijk nog meer ontwikkeld op zowel persoonlijk als inhoudelijk vlak. Je bent begaan met de carrières van je promovendi en daar ben ik je heel dankbaar voor. Ten slotte hebben we van tijd tot tijd enorm kunnen lachen, iets wat misschien nog wel het belangrijkst is!

Wim, je promotietraject startte een aantal maanden voor dat van die mij, waarin jij de "opvolger" van het project van Sander was. Ik heb met veel bewondering gekeken naar de manier hoe je werkt, efficiënt en gestructureerd, waarmee je reuzenstappen maakte. Het was een voorrecht om met jou en Dorien een groot deel van mijn promotietijd de kamer te delen; ik heb enorm veel van jullie kunnen leren. Dank voor alle gezelligheid, levenslessen en hulp tijdens mijn promotie!

Maikel, dit zal je eerste -en waarschijnlijk ook laatste- promotie zijn die je mee gaat maken. Van tijd tot tijd vraag je je af wat ik (buiten koffie drinken) doe, dus bij deze heb je een boek waar je aan kunt beginnen! Daarbij ken ik weinig mensen (buiten mama) die zoveel kunnen praten als jij, dus daar kun je dit ook wel voor gebruiken ;).

Geachte leden van de beoordelingscommissie, bestaande uit prof. dr. Hans-Peter **Brunner-La Rocca,** prof. dr. Tilman M. **Hackeng,** prof. dr. Ron **Kusters**, en prof. dr. Leon J. **de Windt**, dank voor het beoordelen van mijn proefschrift. Prof. dr. Rick **Body,** thank you for the time and effort taken in reviewing my PhD thesis.

Dr. **Joyce** van Beers, dr. **Douwe** de Boer, dr. **Judith** Bons, dr. **Jan** Damoiseaux, prof. dr. ir. **Yvonne** Henskens, dr. **Alma** Mingels, dr. **Irene** Körver-Keularts, dr. ir. **Sander** Streng, dr. **Bart** de Wit, en dr. **Will** Wodzig**,** allen hartelijk dank voor jullie interesse in mijn onderzoek, de kritische vragen en de gezelligheid op de werkvloer maar ook daarbuiten tijdens verschillende borrels. Douwe en Sander, uiteindelijk is het bij jullie ooit begonnen; de afstudeerstage van mijn bachelorstage is me dusdanig bevallen dat dit uiteindelijk het resultaat is, enorm bedankt daarvoor! Professor van Dieijen-Visser, beste **Marja**, dank voor de stevige en inspirerende (troponine) fundering die je mij, alsook de afdeling, gegeven hebt.

Tijdens een promotieonderzoek heb je het voorrecht om met heel veel slimme mensen te

mogen samenwerken; zowel binnen als buiten het laboratorium. De collega's die hier het dichtst bij stonden waren de mede promovendi en onderzoekers.

Noreen, we hebben (helaas) maar kort samengewerkt maar deelde veel gemeenschappelijke interesses. Dank voor alle hulp (ook later in het traject); ik had bewondering voor de energie die je overal inlegde. Ondertussen ben je (helaas) een beetje burgerlijk geworden; ik wens je het allerbeste toe!

Dorien, ik heb enorm veel van je kunnen leren en ik ben heel erg blij dat ik toch een groot deel van de promotietijd met jou een kamer mocht delen. Ook alle (sportieve) uitjes en biertjes naast het werken waren heel leuk! Ik heb er heel veel vertrouwen in dat jij uiteindelijk gaat komen waar je wilt komen! Je moet snel weer eens naar Maastricht komen want anders zijn wij genoodzaakt een bezoekje te brengen aan Klein-Dongen.

Judith, dank voor alle hulp tijdens mijn promotieonderzoek. Het was heel knap hoe je alles hier in Maastricht combineerde met Rotterdam, en ik weet zeker dat je een heel goede dokter gaat worden! Ook dank voor het verzorgen van mijn post-puberale opvoeding samen met Dorien en Wim.

Stephanie, dank voor alle hulp en gezellige tijd samen! Recent ben je ook gepromoveerd, een hele knappe prestatie gezien de combinatie van het CDL met M4I!

Max, jij startte enige tijd na mij als onderzoeker bij het CDL en we hebben mogen veel tijd mogen delen. Dank voor alle hulp, voor het relativerend vermogen maar vooral voor de enorme bak aan algemene kennis die je bezit; het zou een stuk minder leuk zijn geweest zonder jou!

Ellen, jij bent de 'laatste' aanwinst van het troponine groepje. Dank voor alle hulp aan het einde van mijn promotietraject, en heel veel succes met de rest van je onderzoek! Ondertussen ben je alweer even bezig, en ik vind het heel knap hoe je de verschillende projecten (van celkweek tot kliniek) allemaal managed!

Naast de collega's was het laboratorium ook een plek waarin ik met vele studenten heb mogen samenwerken. **Ralph, Issam, Michelle, Laura, Ruud, Tosca, Lieve, Bauke, Caspar, Cristina, Glenn, Marc, Maurice, Mike, Manouk, Loes, Felicia, Demi, Daphne, Iris,** dank voor de samenwerking en gezelligheid! **Jella** en **Jordy** jullie hebben beiden enorme bergen werk verzet op het hypoxie project; dank voor alle hulp en late uurtjes tijdens het celkweken! Ook beide heel veel succes met jullie promotietraject, ik heb er heel veel vertrouwen in dat het helemaal goed gaat komen! **Anina, Micha, Bram, Floor, Hanna, Eline,** dank voor alle hulp en inspanningen tijdens de studie verschillende studies op de spoedeisende hulp!

Naast alle onderzoekers wil ik alle collega's binnen het CDL bedanken. **Lucie**, **Jessica**, en **Rachelle** bedankt voor het regelen van allerlei zaken. **Maurice** en **Richard**, bedankt voor alle hulp met de eindeloze bestellingen en het gesjouw. **Fritz**, ofwel **Vincent** (welke van de twee

was je echte naam?, bedankt voor alle hulp (en opvoeding) de afgelopen jaren. Je bijnaam is 'de levende lablegende' en dat is niet voor niks! **Jeffrey,** succes met het opvolgen van Vincent, dat gaat helemaal goed komen! **Petal**, bedankt voor de gezelligheid op het lab en tijdens borrels. Ik zal niet meer wegrennen! **Serva** en **Ingrid,** dankjewel voor al de hulp en oplettendheid bij tal van zaken. De collega's van de ICT, in het bijzonder **Jan, Ramon, Annemieke, Alex, Rick,** dank voor het faciliteren van allerlei zaken en het continue bereid zijn om (weer) vragen van me te beantwoorden! Zonder jullie waren veel van de machine learning-gebaseerde projecten niet geworden wat ze nu zijn.

Anne-Marije, het is helaas anders gelopen dan we beiden waarschijnlijk gehoopt hadden. Toch wil ik je bedanken voor de tijd samen, ik heb enorm veel plezier met je gehad en daarbij ook veel van je mogen leren. **Toine,** je hebt me heel veel aan het denken gezet, dank voor dat en ook de gezelligheid!

Prof. Dr. **Martijn** Brouwers en Dr. **Yuri** Foreman, bedankt voor de samenwerking omtrent het glucose stuk. Yuri, dank voor alle hulp en gezelligheid, heel veel succes met je opleiding tot internist!

Dr. **Patricia** Stassen en Drs. **Paul** van Dam, dank voor de vruchtbare samenwerking van de afgelopen jaren welke hopelijk nog lang zal doorgaan. Ik waardeer de uitwisseling met jullie enorm, en hoop dat onze projecten doorslaggevend gaan zijn! Drs. **Lars** Lambriks, dank voor alle en gezelligheid bij het regelen van allerlei zaken bij onze studies. Heel veel succes met het zoeken van een baan; ik heb er alle vertrouwen in! Drs. **Hella** Borggreve, Drs. **Maaike** Schalkwijk, en Drs. **Judith** Stoffers, dank voor de hulp bij de eerste machine learning versus dokter studie!

Dr. **Bas** Bekkers, drs. **Yvonne** van Cauteren, dr. **Martijn** Smulders, en dr. **Jeroen** van Suijlen, bedankt voor jullie wetenschappelijke input bij de verschillende publicaties rondom cardiaal troponine.

Prof. dr. **Leon** Schurgers en Drs. **Niko** Deckers, dank de wetenschappelijk discussies en voor alle hulp bij het synthetiseren van het cardiaal troponine! Beste dr. **Kasper** Rouschop, dank voor de zeer waardevolle feedback op het hypoxie stuk en daarnaast dank voor de mogelijkheid om gebruik te maken van uw hypoxische kamers gedurende het opzetten van de studie! Dr. **Freek** Bouwman, dank voor alle hulp en tips bij het analyseren van de eiwit adducten middels LC-MS/MS.

Beste dr. **Leo** Jacobs en dr. **Floris** Helmich, dank voor jullie hulp bij het multi-center machine learning project. Het was heel fijn dat jullie bereid waren om mee te werken aan onze studie in de vroege fase!

Dear prof. dr. **Tobias** Breidthardt and prof. dr. Christian **Mueller,** thank you for the collaboration on the diurnal variations of natriuretic peptides!

During my PhD I was able to perform a three-month research visit to Clarkson University in Potsdam (New York). Dear prof. dr. **Costel** Darie, prof. dr. **Artem** Melman, it was an honor and pleasure to be part of your scientific groups! Thank you for all the time and effort taken to teach and help me with my scientific projects. Artem, it was with great sadness that I heard about your death, I wish to extend my deepest sympathies with your family. Dear dr. **Vasantha** Kadambar, I owe you most of my thanks. Thanks for the endless help with all the experiments! I wish you best of luck and joy together with your family back in India. Dear **Devika, Madhuri, Cristiana, Emma, Roshi,** thanks all for the help during my time at the biochemistry lab! **Cristiana,** thanks for all the fun we had together and little trips we conducted. Dear **Garegin,** I can't even recall how we met but I am very grateful for the time we spent together. My trip to Potsdam wouldn't be the same without meeting you, good luck in the field of computer science! Dear prof. dr. **Cameron** Koch, "father" of the EF5 compound, thank you for all the help with the hypoxia project. Your knowledge and suggestions significantly contributed to a successful completion of the project!

Vrienden van "thuis-thuis" zijnde **Dirk, Yannick, Noëll, Giel, Wouter, Guus, Joeri, Jelte, Cem, Kim, Malou, Kelly:** ik vind het mooi om (bijna) elk weekend nog in het vertrouwde Heesch te zijn. Op naar nog vele jaren! **Bart** en **Joost,** het is leuk dat we naar al die jaren naar de HAVO nog contact hebben en regelmatig fietsen. Lieve **Kim,** het is gek dat we elkaar pas na zoveel jaar leerde kennen in Maastricht. We hebben het heel fijn samen en ik hoop dat dat nog heel lang mag duren!

Talenten van cohort 11/12, zijnde **Jos, Joep en Eric**. Inmiddels over het hele land verspreid met ieder zijn eigen weg en mening, over één ding zijn we het eens: HBO was de mooiste tijd uit ons leven. Ik vind het mooi dat we elkaar nog een aantal keer per jaar zien.

De oud-huisgenoten van 'Amby', zijnde **Marsha, Helena, Thijs, Wybren** (en eigenlijk ook **Koen**), we hebben slechts één jaar samengewoond maar zien elkaar nog regelmatig. Dank voor alle gezelligheid en jullie altijd ongezouten mening!

Mannen van Viramitas, een fantastisch sportieve groep binnen Maastricht. **Aaron**, **Jeroen,** succes met het afronden van jullie promoties. **Onno, Timo, Wybren, Rafael, Antal, Paul, Stijn, Coen,** dank voor alle gezelligheid tijdens en na de sportieve activiteiten. **Felix, Chris, Roeland, Huub, Jochem, Luc, Sjoerd, Sem, Tristan, Bob,** dank voor alle gezelligheid tijdens jullie tijd in Maastricht. ****maten, zijnde **Frank, Bram, Joost** en **Sjors**, ooit begonnen op een dinsdagavond, verdergegaan als ontbijt en inmiddels een waar begrip in Maastricht. Meesterlijk! **Frank,** nog een speciaal woord aan dank voor jou gezien al je werk voor m'n boekje!

Lieve **mama** en **papa**, op alle mogelijke manieren hebben jullie ervoor gezorgd dat ik hier nu sta. Ondanks dat het inhoudelijk vaak lastig was om te begrijpen waar mijn onderzoek precies over ging, toonde jullie altijd interesse en betrokkenheid. Dankjewel voor alle onvoorwaardelijk steun en liefde!