Gupta Strategists

The World is a Clinical Trial

Leveraging existing clinical data to improve healthcare, based on a novel fingerprinting algorithm

Amphia

Gupta Strategists

The World is a Clinical Trial

Leveraging existing clinical data to improve healthcare, based on a novel fingerprinting algorithm

Part of the Gupta Strategists series of studies on practical applications for Big Data

April 2017

Table of contents

Executive summary

Medical decision making is becoming ever more difficult. To make the best possible decisions, all relevant supporting information - both from evidence and from experience - needs to be easily accessible and structured so that it is relevant and accurately applicable to the specific patient. We have tools at our disposal to incorporate evidence into the medical decision making process. Tools to systematically learn from experience - both one's experience as well as that of colleagues - have not yet come to fruition. We need to access, intelligently structure and filter this data for relevance. And we could well use some automation.

This paper presents an algorithm that allows doctors to systematically incorporate their individual as well as collective experience into the decision-making process. The algorithm was developed and tested on data from Amphia hospital in Breda, The Netherlands. This hospital uses the EPIC EHR and has a HIMMS stage 6 classification, meaning it meets high standards on administration and use of data. Our algorithm consists of three steps:

1. Create a unique fingerprint for every patient. As inputs we used: reason for admission, previous diagnoses, medication, lab results, prior surgeries and other treatments.
2. Compare similarities between fingerprints. The algorithm can quickly sort through many thousands of other fingerprints to find all highly similar fingerprints.
3. Develop recommendations. For fingerprints that have the highest similarity with the subject, we provide feedback based on treatment decisions of these highly similar patients. This allows the admitting physician to learn from the experience all other physicians have tacitly accumulated by treating highly similar patients.

As a proof-of-concept, we have shown that the fingerprint model can explain and predict two clinical outcome measures: 1) length of stay and 2) probability of ICU admission. The fingerprint model matches or surpasses the best existing and published models. The algorithm can be applied to other outcome measures as well, and can provide decision recommendations based on outcome differences within highly similar patient groups.

We see three possible areas of application for the algorithm:

- *Learning on the go:* an upgraded version of the algorithm in its current state. We focus exclusively on the 'live' data recorded daily by care professionals. We do not take into account, or attempt to reconcile these data points, with the 'formal' publications; combining this experience- and evidence data points would be a useful next step.
- *The world is a clinical trial:* a new way of performing clinical research, that can deal with much greater variance in patient characteristics than typical randomized controlled clinical trials. Results are applied in practice immediately. The predictive power of the model evolves with the population and the care regimes. In that sense, it is a 'self-learning' model.
- *Management information:* the algorithm can be applied for payors and providers alike, for example in the prediction of cost and scheduling.

Bio Amphia

Amphia is one of the largest general hospitals in the Netherlands and one of the 28 top clinical teaching hospitals. Our core tasks, apart from top medical care, are training and research. The hospital has two branches in Breda and one in both Etten-Leur and Oosterhout.

Because of its extensive range of specialist medical treatment services and its solid position in the region, Amphia has a large care area (over 425,000 inhabitants). Among the treatment services, 33 different speciality fields are represented. The Hospital focuses mainly on oncology, heart, vascular and lung diseases, exercise and movement of the body, women, mother & child and healthy ageing.

Over 270 medical specialists, 4,300 members of staff and 340 volunteers work on a daily basis to carefully deliver optimal quality care in a safe environment. With an eye for our patients and the people that are important to them.

For more information, go to www.amphia.nl

Introduction

Medical decision making is becoming ever more difficult

The volume of medical research is growing exponentially[1]. It is impossible for a single physician to remain abreast of all, or even the most, important developments in their field. Not only are the demands of everyday clinical practice too intense to allow much time for study: the volume of research and the speed at which it grows is such that even full-time devotion is not sufficient to process all research driven information being produced. This particularly holds true for all the clinical decision based information being generated continuously around the world, which is the subject of this study.

Real life clinical decision support can potentially make medicine better. But at the same time, quality of analyses and decision making may be compromised due to information overload. With the growing amount of available information, the act of making decisions has become more difficult - not less. It is simply impossible to rationally weigh and judge all information required to make the best decision for the patient without some help and simplification. We need filtering for relevancy; we need intelligent comparisons; and we could well do with some automation.

Evidence based decision support systems already exist; we present a novel decision support tool that is based on doctors' experience

To make the best possible decisions, all relevant supporting information - both from *evidence*[2] and from *experience*[3] - needs to be easily accessible and structured so that it is relevant and accurately applicable to a specific patient.

We have tools at our disposal to incorporate *evidence* into the medical decision making process. Medical publications are available in central repositories over the web and are also searchable, though with some serious limitations (see our study on medical literature, "On Benches and Beds")[4]. Over the past decade, several so-called clinical decision support systems (CDSS) have been developed to make the wealth of information easily available at the point of decision. The appendix provides more background on these systems. CDSS's typically rely heavily on evidence.

[1] US National Library of Medicine, which keeps track of the number of new Medline / Pubmed indexed papers

[2] By evidence, we mean well documented, hard scientific evidence. This source is especially wealthy, with 12.000 articles being added to Medline every week. This source, however, has a drawback. It describes which treatment is best for a comparable group of patients: patients with the same weight, with the exact same diagnosis, with the same clinical history, and combinations of patient characteristics. This exact combination of restrictive traits is very unlikely to present itself to the specific patient a doctor is dealing with in real life. That is where experience enters the equation.

[3] Experience is a large category of everything that isn't, strictly speaking, hard evidence. Rather, it is complementary to it: it is the translation of what is known to be best in a controlled environment to what the doctor(s) believes to be the best treatment in an uncertain world. On the one hand, this may be a hospital protocol that specialists have devised to define how they apply science and their own expertise to patients. On the other hand, it is the doctors' individually acquired and exercised experience. This is the collective set of all observations and decisions.

[4] Kruif TM de, Laar L van de, Hagenaars, N. 'Maat en getal bij de biomedische onderzoeksagenda van Nederland'. Ned Tijdschr Geneeskd. 2017

In addition to ‘formal’ reviewed or non-reviewed publications, over the past decades, health care organizations have amassed a vast amount of electronic data on clinical parameters, treatment decisions and outcome. However, such data is typically applicable only to specific groups and specific decisions or interventions - and much of what doctors do on a day-to-day basis is based on an amalgam of their own knowledge and *experience*, albeit having grown from the evidence. And since every doctor does this individually, and presumably in a different manner, there is huge value present in a doctor’s mind and sources in which doctors record their decisions, such as EHRs.

Tools to systematically learn from *experience* - both one’s experience as well as that of colleagues - have not yet come to fruition. Certainly, plenty of data is being collected - from sources such as electronic health records (EHRs), automated hospital pharmacy prescription systems and insurance claim administration. Often, this data has initially been collected for special purposes with a narrow focus. Occasionally, it has been collected with no clear sense of purpose at all. But as far as we know, such data has not yet been used to enable doctors to systematically incorporate learnings from their own experience, as well as that from their colleagues, into their decision-making process.

Many of the poster-child techniques of ‘big data’ processing were developed for medical and life-science purposes: the challenges of medical image processing and genome sequencing have been key drivers of the machine learning revolution. When it comes to making decisions for individual patients’ health care, however, clinicians are still mostly left to their own preferences. Their access to the collective knowledge is limited by the time they can afford to keep up with research or consultation with colleagues, or the inclination to do so. This variability in preference is itself a source of contention, one we will not be addressing here.

Our point of departure for this study was the premise that a concept that makes the collective clinical experience data points constantly being created and recorded around the world available at the point-of-care would be useful. A mathematically well-defined learning algorithm based on a collective set of experiences could help address, if not eliminate, an important source of unwanted variation in medical practice. And it could help identify and thus accelerate uptake of best practices from other clinicians.

This paper presents an algorithm that allows doctors to systematically incorporate their individual as well as collective experience into the decision-making process

In this paper, we explore whether the data that hospitals already collect and store can be used to improve health care outcomes for patients, and explore ways that this information can be leveraged to the benefit of patients, hospital managers, insurance companies, and policy makers. We focus exclusively on the 'live' data recorded daily by care professionals. We do not consider, or attempt to reconcile these data points, with the 'formal' publications, though undoubtedly combining the two would be a useful next step.

Even from a computer science point of view, medical decision making is a particularly hard problem: for every patient characteristic, a doctor needs to weigh all possible scientific evidence and all data points in one's own experience. More formally: if there are n evidence data points, p patient characteristics and m experience data points, the problem is of order O*(m*n*p)* - i.e., 3-dimensional. Since *m, n* and *p* are large and additionally, *m* and *n* grow very quickly, the problem becomes more and more time consuming to solve. In practice all of this is of course undertaken subconsciously by the clinician and not explicitly as a brute force mathematical approach. And it is this subjectivity that makes it hard to trace the source of observed variations in clinical practice, let alone judge the validity of the observed variations.

We present an algorithm which analyzes a collective set of experiences by reducing two dimensions (patient characteristics and experience) to a single-value, thereby greatly reducing the complexity of inferring conclusions from all relevant data points. This algorithm extracts, analyzes and re-presents the information in medical practice as a support for the individual medical decision making. Hence the title of this study: 'The World is a Clinical Trial'. We take the view in this study that every single clinical decision doctors make every day is a new data point, that can be reviewed and considered for relevance in a new situation. In this sense the clinical world as we know it, is itself an ongoing trial, that we can leverage to improve medical outcomes. To the best of our knowledge, this is the first model to aid decision making based on experience. The third dimension (evidence) is not explicitly included in this algorithm. In the discussion section, we describe how evidence can also be taken into account explicitly; this is a relatively easy step.

Our aim is not to replace the individual doctor's decision making with an algorithm - the vagaries of medical practice require human intervention. Our aim is much more modest - we propose to support human decision making with an algorithm that systematically accounts for how other doctors decide in similar conditions.

Methodology

Using fingerprints to incorporate experience into decision making

Fingerprints – the concept in three steps

We hypothesized that based on data from a good Electronic Health Record (EHR), in which data is organized in a structured fashion, it is possible to create highly distinctive 'fingerprints' for each patient. A patient specific 'fingerprint' is composed of clinical variables such as lab values, current and past reasons for admission, previous therapeutic interventions, etc. Moreover, fingerprints could contain an element of time-decay: abnormal lab values recorded last week could weigh more strongly in the fingerprint than those recorded a year ago. For this study we focused on unique fingerprints for elderly patients (70+) at the time of admission.

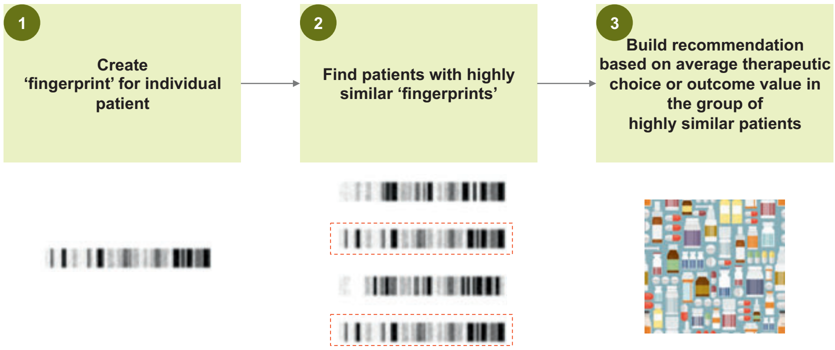
Once fingerprints were created for individual patients at each time of admission, we built an algorithm that can quickly sort through many thousands of other fingerprints to find all highly similar fingerprints - akin to such lab techniques as gel electrophoresis which is used to, for example, analyze DNA patterns (see figure 1 below).

Finally, once a group of highly similar patients is identified, we can determine a 'recommendation' for the current patient based on the average values observed in the comparison group. We could, for example, predict length-of-stay, determine the chance of ICU admission or give a recommendation on which, if any, antibiotic to prescribe.
This allows the admitting physician to learn from the experience all other physicians have tacitly accumulated by treating highly similar patients.

Figure 1 shows an overview of the concept based on which this study was built.

FIGURE 1:

The concept of 'fingerprinting'


1
Create 'fingerprint' for individual patient
2
Find patients with highly similar 'fingerprints'
3
Build recommendation based on average therapeutic choice or outcome value in the group of highly similar patients

More than simply proposing a theoretical concept, the purpose of this study was to build a proof-of-concept model for the fingerprinting algorithm, and to study how well it can predict certain clinical outcomes or recommendation therapeutic choices. The model should be fast enough so that it works in real-time within an actual EHR and thus function as a full-blown CDS that builds on all experience contained in its database.

Constraints / limitations to this study

As we will describe below, our focus was on developing an algorithm that could put this idea to practice. We wished to test, to the extreme, how much predictive information is embedded in the data itself. As such, we put a few, admittedly artificial, constraints on our design - which would of course not apply in real life implementations of this algorithm (see *discussion* section for more detail):

- Let the 'experience' data speak for itself. The idea focuses on 'experience' data points and does not include any 'evidence' based validation. We wanted to see just how much relevant information is contained in raw data, and add as little clinical interpretation as possible (combining evidence or protocols is one way to enrich the raw EHR data).
- As such, variables in fingerprints are not weighted in any intelligent way. Certainly, in real-world applications of this idea, it is conceivable that intelligent weighting (based on clinical evidence, or machine learning) would add considerably to the predictive power of the model.
- Limit data manipulation to an absolute minimum. We wanted the model to work with raw EHR data as closely as possible, so that the model could be directly implemented in actual EHRs without building separate datasets. For example, we

chose not to create artificial value groupings for lab values, but rather used their absolute values. The only concession we made on this point was that we made some minor manipulations to open text fields (such as 'reason for admission') in order to categorize them, which can be done quickly on-the-fly.

Data set

To perform this study, we were given access to an anonymized extract of EHR data from Amphia hospital in Breda, The Netherlands. Amphia hospital uses the Epic EHR system, and was one of the first Dutch hospitals to receive HIMSS Stage 6 qualification meaning its EHR and clinical processes meet very high, externally audited standards. As such, it formed a very suitable basis on which to build this proof-of-concept model.

The contents of the dataset available for this study can be summarized as follows:

- all admissions of patients age 70+ from January 2014 to April 2015
- all clinical data points stored in the EHR from the following six categories:
    - o reason for admission (current and past) - 294 unique values
    - o previous diagnoses - 1,474 unique values
    - o prior surgeries - 644 unique values
    - o prior other clinical interventions - 215 unique values
    - o lab values - 254 unique values
    - o medication (by ATC5[5] code) - 377 unique values
- historical data from January 2013 onwards, so that for all admissions there is at least 1 year of prior data available in the dataset

In total, the dataset contains information on 35,894 admissions. Table 1 summarizes the demographic characteristics of the patients in the dataset.

[5] ATC5: first 5 digits of the Anatomical Chemical (ATC) code, according to the classification maintained by the World Health Organization (WHO). Represents chemical/therapeutic/pharmacological subgroup

	Total	Test set	Reference set
Total # of admissions	35,894	6,988	28,906
Total # of unique patients	18,927	3,785	15,142
Age categories at admission			
-70-74 year	28%	26%	29%
-75-79	30%	32%	29%
-80-84	23%	23%	23%
-85-89	13%	13%	13%
-90+	6%	6%	6%
Percentage female	52.5%	52.8%	52.4%

Table 1: characteristics of the dataset

Test set vs. reference set

To develop and validate the model, and to eliminate the risk of prediction bias, we created a test and reference set of patients. The test set of patients represent a group of patients admitted to the hospital for whom we want to make predictions, and a reference set of patients are those whose data points can be used to make predictions for the test set. These sets were constructed as follows:

- The test set contains 20% of all unique patients from the dataset. For these patients, we selected all admissions in the 2nd half of 2014. This selection ensured that 1) the tests we apply would be 'realistic', in that we would only draw conclusions from analyses of historic data in the reference set, and 2) there would be enough data in the dataset to accurately measure such outcome variables as length-of-stay. In total, the test set contains data on 6,988 admissions for 3,785 unique patients.
- The reference set contains the remaining 80% of all unique patients from the dataset. For these patients, we selected all admissions in the 1st half of 2014. In total, the reference set contains data on 28,906 admissions for 15,142 unique patients.

Table 1 summarizes the contents of both datasets.

Fingerprints by category

For every unique admission, we created a fingerprint for each of the six data categories available to us. To do so, we used the following procedure:

- Encode all variables
    - o Categorical variables were converted to Boolean variables. For example, for the variable 'previous diagnoses', all possible diagnoses were converted to distinct variables, and each variable set to either 0 (diagnosis not present) or 1 (diagnosis present). This applies to all variables except lab values, which are numerical. For medication, we only used data on which medication a patient used, not in what quantity, frequency or dosage.
    - o Lab values were converted to a 'distance from normal'. For all lab values, a 'normal range' is recorded in the EHR based on the population and particular lab conditions at the time of measurement. 'Distance from normal' for the purpose of this study was then defined such that a value of 0 represents the precise median between the lower and upper value of the normal range, a value of -1 represents the lower value of the normal range, and +1 represents the higher value of the normal range.
- Time code variables
    - o All variables were linearly weighted in time such that:
        - ▪ Values up to 1 year prior to admission date receive a weight of 0.
        - ▪ Values at date of admission receive a weight of 1.
        - ▪ All remaining values receive a weight between 0 and 1, with the weight linearly decaying from 1 at admission to 0 one year post admission.
        - ▪ Create vector for each of six categories: For each of six categories, we calculated a multidimensional vector based on all the values within that category. This vector is the mathematical representation of a 'fingerprint'.

Fingerprint density

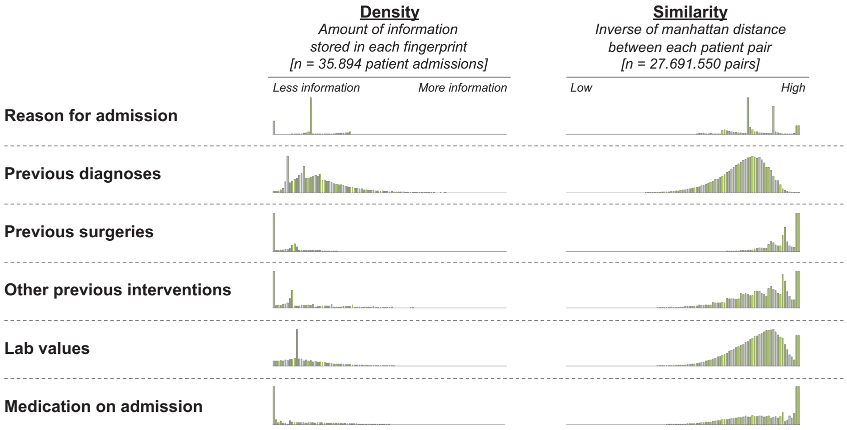
It is possible for a patient to have a fingerprint without any previous data points. Of course, such a fingerprint is much less distinctive than a fingerprint for an admission where we have many data points. Therefore, we wanted to create a measure for how much information is contained in each fingerprint. We call this measure the 'density' of a fingerprint. To create this measure, we used the following procedure for each of the six information categories:

- Determine maximum vector length, defined as the length of the hypothetical vector in which all variables are set to the maximum observed value
- Density = (observed vector length) divided by (maximum vector length)

As can be inferred from the above methodology, density values for each category can range from 0 to 1, where 0 is minimum information density and 1 equals maximum information density. Figure 2 provides frequency distribution graphs for densities within each of the six information categories.

FIGURE 2:

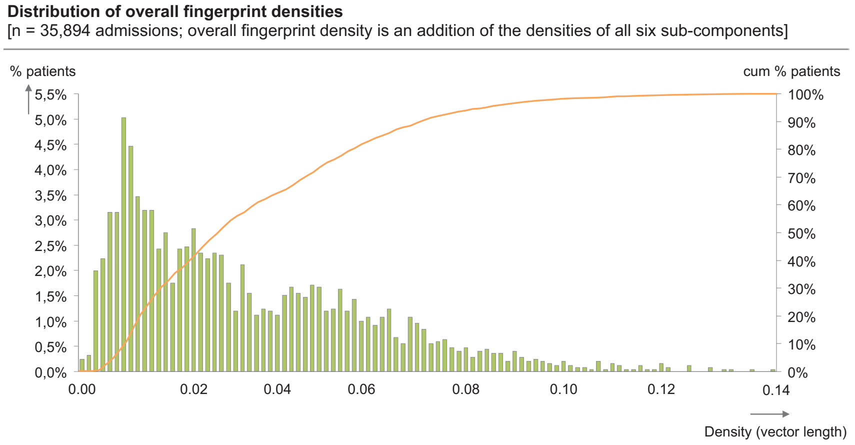
Distribution of densities and similarities for individual fingerprint components (Note: for an explanation of 'manhatta n distance', please refer to the next section on similarity score calculations)



To determine an overall density for an admission, i.e. based on all six data categories combined, we add the densities of each of the six categories. Figure 3 provides a frequency distribution graph for overall fingerprint densities of all 35,894 admissions available in the dataset.

FIGURE 3:

Distribution of overall fingerprint densities

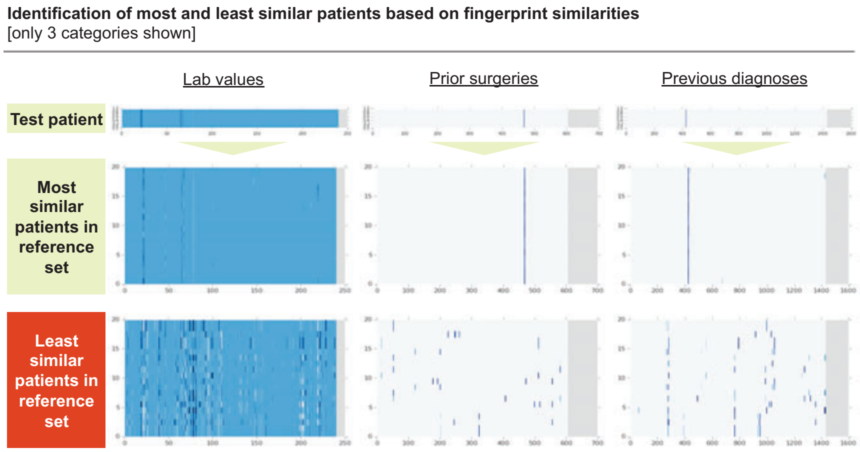


Similarity

We can determine similarities between each of the 6,988 admissions in the test set and each of the 28,906 admissions in the reference set, for a total of 27.691.550 pair-wise comparisons. Figure 4 provides a graphical representation of the process of comparing fingerprints between the test and reference sets.

FIGURE 4:

Graphical representation of the process of fingerprint comparison


Identification of most and least similar patients based on fingerprint similarities
[only 3 categories shown]
Lab values
Prior surgeries
Previous diagnoses
Test patient
Most similar patients in reference set
Least similar patients in reference set

Similarities are calculated for each of the 6 information categories in the following fashion:

- Determine the 'Manhattan distance' between vector of the test subject and vector of the reference subject. Manhattan distance is a very simple method of calculating distances between 2 points in a multidimensional vector system. The method's name reveals how it works: it can be used to determine how many 'blocks' one must travel to get from, e.g., 1st avenue and 21st street, to 2nd avenue and 23rd street in Manhattan. In this case, the Manhattan distance would be 3. While in the case of Manhattan, there are only 2 dimensions (avenues and streets), the calculation easily expands to vectors of many dimensions.

- Divide each distance by the maximum potential distance for that information category - i.e., the distance between a vector where all values are 1 and a vector where all values are 0. This division creates a distance that is always between 0 and 1, where 1 represents maximum potential distance and 0 represents 100% equality.
- We now have a distance between 2 vectors, but we find similarities conceptually more intuitive to work with. We convert using the following formula: similarity = 1.0 - distance. Now, a similarity value of 1 means 100% equality between 2 vectors.

Figure 2 above displays, for each data category, frequency distributions of similarities between 6,988 test-set admissions and 28,906 reference-set admissions, for a total of 27.691.550 pairs.

To calculate similarities between 2 admissions based on multiple data categories (e.g., based on both prior diagnoses and lab values), we multiply similarities for each of the categories. For example:

- 100% similarity on prior diagnoses and 100% similarity on lab values translates to an aggregate similarity score of 1 x 1 = 1
- 100% similarity on prior diagnoses and 50% similarity on lab values translates to an aggregate similarity score of 1 x 0.5 = 0.5
- 50% similarity on prior diagnoses and 50% similarity on lab values translates to an aggregate similarity score of 0.5 x 0.5 = 0.25

Note that in theory, this method puts a relatively large penalty on dissimilarity. Two patients who are completely similar on 5 categories but completely dissimilar on 1 category would receive a similarity score of 0. In practice, however, this does not occur because 0 values in the fingerprint are also information points. For example, to be 100% dissimilar, a test patient would have to have undergone all possible surgical operations that a reference patient has not undergone, and vice versa.

Note also that the choice to first calculate similarities per category with an outcome between 0 and 1 was a design choice made to ensure that categories for which the vector consists of many components do not dominate the similarity calculation. As we explain in more detail in the discussion section, these design choices could perhaps be improved for real-world implementations.

FIGURE 5:

Distribution of overall similarities between all patient pairs

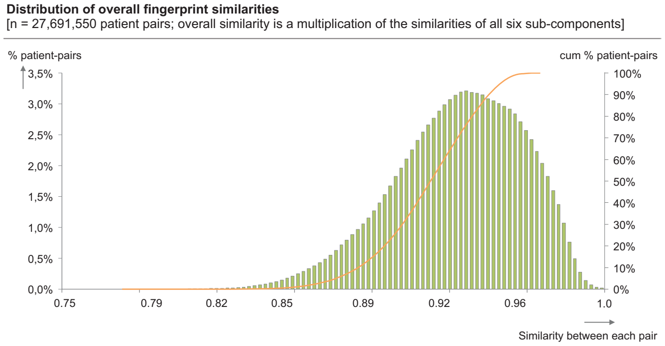


Figure 5 displays aggregate similarities for all 27.691.550 pairs based on all six data categories.

Output variables for comparison

At this point, we have 6 separate fingerprints for each patient, namely, one for each of the six variable categories and a combined fingerprint for all 6 categories per patient. We also have for each pair of patients a similarity score: again, a combined score as well as a similarity score per each category. To validate the developed approach, we then used the test patient set as a proxy for patients that present themselves to the hospital - these are the patients that form our 'trial population'. For this test set we are interested in making recommendations on treatment advice and outcomes based on the similarity in fingerprints predictive system we have developed. And we have for this test set of patients the actual choices made by the clinicians as a measure to benchmark the quality of our predictions[6]. We use the reference set as a proxy for data on which predictions or recommendations would be based.

To test how much clinically relevant information can be obtained from fingerprint comparison, we analyzed two outcome variables:

- **Length of stay:** the number of days between date of admission and date of discharge.

[6] Here we should draw attention to a paradox in our model: we are using the actual choices made by the clinicians as the benchmark to gauge the success of our algorithm and yet the aim of our algorithm is to improve the quality of these very choices. As we move ahead with further development and refining it is conceivable that expert vetting of these choices to make a 'best practice' learning set may be useful. But in the proposed application this is not necessary. We are using the test set merely as a check on the validity of this approach. The application does not purport to predict a 'right' answer but to make available choices made by others in the most similar and thus presumably relevant cases

- **IC admission:** a Boolean indicating whether a patient was admitted to the ICU during this admission

The method of comparison we used for these variables is as follows:

- First, we analyze the fingerprint for a random test subject
- Then, we identify all reference subjects for whom the similarity scores are higher than a certain threshold. Of course, we do not know a prior what the threshold should be for optimal predictions. Therefore, we tested different thresholds - sensitivity tests.
- Lastly, we obtain the average outcome value in the 'best match' reference group for the test subject under analyses.

Findings: The fingerprint model in action

If the model works as hypothesized, we should be able to draw relevant conclusions on clinical outcomes based solely on the fingerprint of the individual being admitted to the hospital. This is done as described in the previous chapter, by comparing that fingerprint to other fingerprints in the database and interpreting the characteristics of patients with highly similar fingerprints. Furthermore, the predictive power of the model should improve as we set higher similarity thresholds or, in other words, limit comparisons to sets with higher similarity.

In this chapter, we will explore how well the fingerprint model explains two clinical outcome measures: 1) length of stay and 2) probability of ICU admission. We believe that other outcomes can be predicted as well. We discuss other options and areas of application in the discussion section.

Length of stay

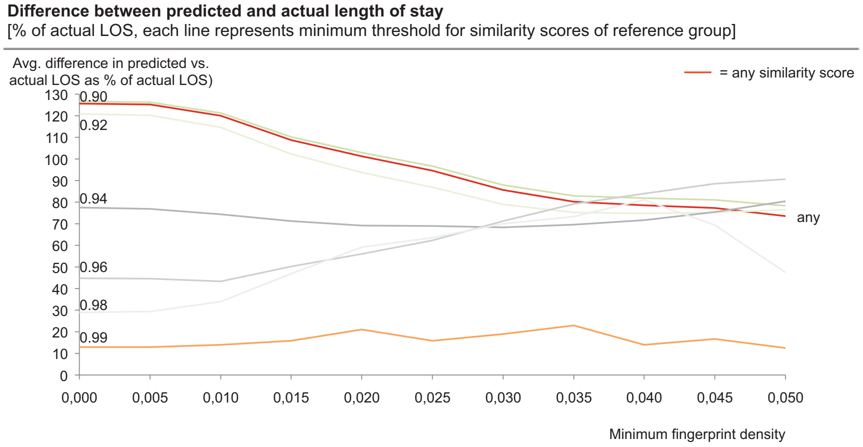
We have attempted to predict the precise length of stay based on an individuals' fingerprint data. Prediction of length of stay is interesting because it facilitates management of bed capacity, it may aid in optimizing use of short-term stay wards and it may help to better prepare and organize the transfer to the patients home or a nursing home (for example, special beds or mobility aids can be requested in preparation of the discharge of patients).

For this analysis, we calculated the average length of stay amongst matching patients from the reference group. We used the difference between 'predicted length of stay' and 'actual length of stay' as the outcome variable. A difference of 0 means that the prediction precisely matches the actual length of stay, and is therefore the best possible model result.

Since we do not know 'optimal' levels for similarity, nor do we know what is the minimum fingerprint density, we ran the model at various thresholds for both variables. Figure 6 shows how well the fingerprint model predicts length of stay at different thresholds for similarity scores, and at different levels for minimum fingerprint density.

FIGURE 6:

Difference between predicted and actual length of stay (LOS) at different thresholds for similarity and density



In this analysis, only patients who meet the minimum fingerprint density level and similarity levels are included. Therefore, the results tell us how well the model can predict the length of stay, provided the threshold values are met.

Overall, we can draw three important conclusions:

- As expected, as we set higher thresholds for similarity scores, meaning that the reference group becomes more similar to the test subject, the predictive power improves significantly
- At very high thresholds (> 0.98 similarity score), the model functions extremely well: it can predict length of stay on average within less than 1 day from the actual length of stay.
- The predictive power of the model declines slightly as we become more restrictive

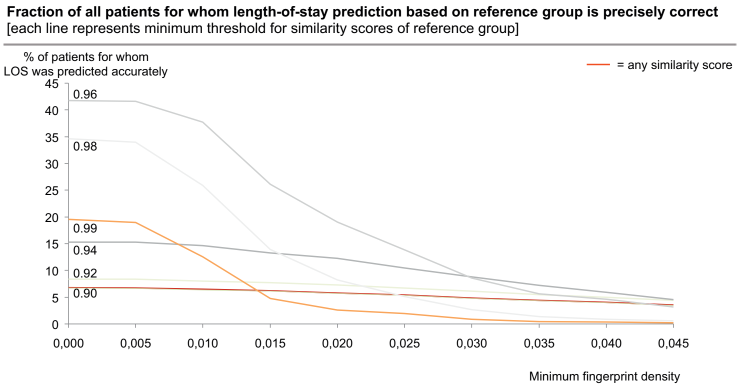
on the minimum fingerprint density of the test subject, meaning that we require more data points for the subject. One would perhaps expect the opposite since higher density implies more comprehensive match with the reference set of multiple vectors. However, here we must consider the second effect of increasing thresholds: the sample size decreases with increasing match requirement. As the threshold fingerprint density increases the number of reference subjects declines rapidly. Requirements of matching on more data points implies fewer patients are included in the analysis, and it becomes increasingly difficult to find matches at any given similarity threshold. For example, at a minimum similarity threshold of 0.98, we find an average of 371 matches if we set the minimum fingerprint density at 0.01. In contrast, we find only an average of 4 matches if we set minimum density at 0.04. We expect this effect can be reduced quite easily by using larger datasets.

We also looked at length of stay prediction from a different perspective. We wanted to know for what fraction of patients the model can predict the length of stay (in number of days) precisely correct.

As in the previous analysis, we ran the model at different levels for minimum similarity and minimum fingerprint density. Figure 7 shows the results of these analyses.

FIGURE 7:

Percentage of precisely correct length-of-stay predictions (LOS) at different similarity and density thresholds

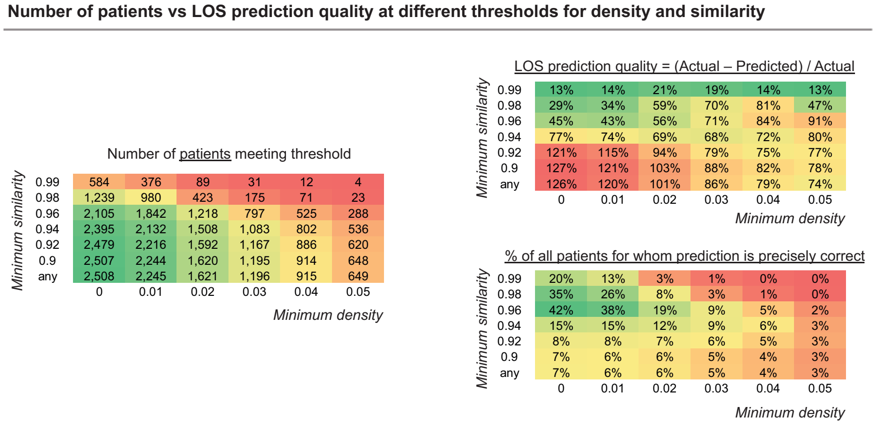


Here also, we see that predictive power of the model becomes better as we set higher thresholds for similarity. However, we also see that the quality of the model diminishes at very high similarity thresholds (> 0.96). This, again, is likely due to the fact that we find ever fewer matches at such high similarity levels. We also see that overall usefulness of the model diminishes quickly as we increase the fingerprint density thresholds. This is because fewer patients are included in the analysis, so that the group for which the model can make no prediction also grows rapidly. Figure 8 shows a more detailed view of the relationship between density and similarity thresholds, number of patients meeting those thresholds and quality of LOS prediction.

Overall, the results are very encouraging. If we choose the right thresholds for similarity and density, we find that the model can predict length of stay *precisely correct* for up to 42% of patients and predict length of stay within a 13% margin of error on the basis of the clinical information available at the point of admission alone within a single hospital EHR without recourse to any other literature or data. Since such models can be directly linked to the available EHR software at the doctor's desk, they can be quickly integrated into clinical practice helping hospitals manage their clinical capacity better.

FIGURE 8:

Relationship between density and similarity thresholds, number of wpatients and quality of LOS prediction



Number of patients vs LOS prediction quality at different thresholds for density and similarity

Number of patients meeting threshold

Minimum similarity	0	0.01	0.02	0.03	0.04	0.05
0.99	584	376	89	31	12	4
0.98	1,239	980	423	175	71	23
0.96	2,105	1,842	1,218	797	525	288
0.94	2,395	2,132	1,508	1,083	802	536
0.92	2,479	2,216	1,592	1,167	886	620
0.9	2,507	2,244	1,620	1,195	914	648
any	2,508	2,245	1,621	1,196	915	649

Minimum density

LOS prediction quality = (Actual – Predicted) / Actual

Minimum similarity	0	0.01	0.02	0.03	0.04	0.05
0.99	13%	14%	21%	19%	14%	13%
0.98	29%	34%	59%	70%	81%	47%
0.96	45%	43%	56%	71%	84%	91%
0.94	77%	74%	69%	68%	72%	80%
0.92	121%	115%	94%	79%	75%	77%
0.9	127%	121%	103%	88%	82%	78%
any	126%	120%	101%	86%	79%	74%

Minimum density

% of all patients for whom prediction is precisely correct

Minimum similarity	0	0.01	0.02	0.03	0.04	0.05
0.99	20%	13%	3%	1%	0%	0%
0.98	35%	26%	8%	3%	1%	0%
0.96	42%	38%	19%	9%	5%	2%
0.94	15%	15%	12%	9%	6%	3%
0.92	8%	8%	7%	6%	5%	3%
0.9	7%	6%	6%	5%	4%	3%
any	7%	6%	6%	5%	4%	3%

Minimum density

To put these results in perspective, we searched medical literature for other models that predict length of stay. Such models typically focus on particular subgroups, for example primary total knee replacement (Carter et al.[7]) or cardiac patients (Hachesu et al.[8]). Such subgroups are likely to have a more homogenous distribution for length of stay, making it somewhat easier to predict length of stay.

[7] Carter et al., Predicting length of stay from an electronic patient record system: a primary total knee replacement example. BMC Med. Informatics and Decision Making 2014 14:26

[8] Hachesu et al., Use of Data Mining Techniques to Determine and Predict Length of Stay of Cardiac Patients. Health Inform. Res. 2013 Jun; 19(2): 121-129

Moreover, all of these models stratify the outcome variable. That is to say, they do not predict a precise number of days, but rather predict an outcome category (e.g.: 0-1 days, 2-4 days, 5-7 days and longer). Of course, a category prediction may also be useful enough for most applications. If we also stratify our model predictions the quality of the output will likely improve.

As such, it is difficult to compare our results to many other studies. However, our results do appear to compare very favorably. Dent et al.[9] found that they could predict the correct length of stay *category* for up to 35% of emergency department patients. In the study by Carter et al., the model predicts length of stay within 1 day accuracy in only ~30% of cases. In our model, we can predict length of stay for up to 43% of patients precisely correct.

Intensive Care admission

As a second proof of concept we used the fingerprint model to predict admission to the Intensive Care Unit (ICU) at any point during the admission. Such knowledge at the time of admission to the hospital would provide the physician the chance to reroute the patient to other hospitals if the local ICU does not have sufficient capacity - even if the patient is presently not an obvious ICU candidate. It would also aide hospital administrators in capacity management. Finally, ICU admission probability can serve as a proxy for complexity or risk and help the care providers adjust their treatment to meet the predicted risk.

We used the fraction of patients in the reference group that was admitted to the ICU at any point during their admission as the target variable. A value of 0 therefore means that 0% of the reference group was admitted to the ICU, a value of 0.5 means that 50% of the reference group was admitted to the ICU, and so forth.

Of course, we would like to translate such a percentage into a straight 'yes' or 'no' prediction. However, it is impossible to set an 'optimum' threshold. The higher it is set, the more likely we will find true positives, but we will also find more false positives. To deal with this problem, we used the receiver operating characteristic (ROC) method to analyze the quality of the outcome variable.

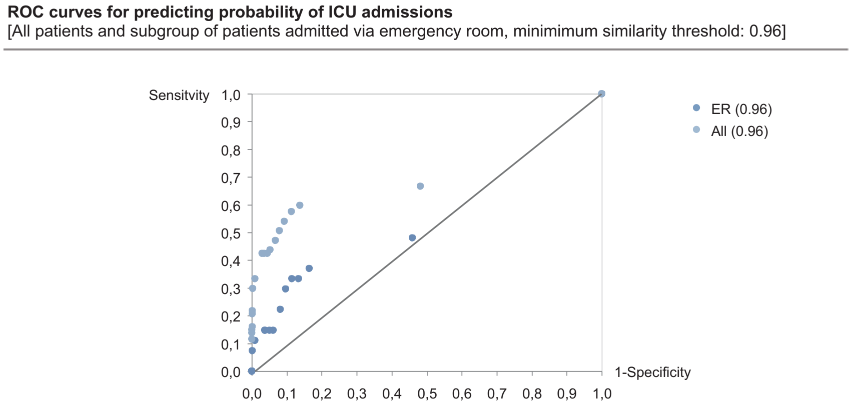
[9] Dent et al. Can medical admission and length of stay be accurately predicted by emergency staff, patients or relatives? Aust Health Rev. 2007 Nov;31(4):633-41.

This technique does not require us to set a threshold value for the outcome variable, but rather, it creates a curve with sensitivity (probability of finding a true positive) on the y axis, and 1-specificity (i.e., probability that the test is false positive) on the x axis. It does so at various threshold levels.

A test that is extremely poor will lead to a 45 degrees straight-line ROC curve. The more the curve bends towards the upper left corner, the better the test discriminates true positives from false positives. The overall quality of the test can be measured by calculating the area under the curve (AUC). An AUC of 0.5 means the test is fully non-discriminatory. While an (hypothetical) AUC of 1.0 corresponds to a test that is perfectly discriminatory.

FIGURE 9:

ROC curves for predicting probability of ICU admissions using the fingerprint model



ROC curves for predicting probability of ICU admissions
[All patients and subgroup of patients admitted via emergency room, minimimum similarity threshold: 0.96]

As in other analyses, we ran the model at various thresholds for similarity scores and fingerprint densities. However, to simplify outputs we do not show all results here, but only show the best results, i.e. at a similarity score threshold of 0.96, and no minimum value for fingerprint density. We ran the model first for all admissions to the hospital, and subsequently also for the subgroup of patients admitted via the emergency room (ER). The results are shown in figure 9.

As can be seen in the graph, the predictive ICU admission rate model for all patients functions remarkably well. At the best threshold value for similarity and density, the test can predict ~60% of ICU admissions correctly, if we accept a false positive rate of ~15% (versus a random-test prediction baseline of 15%[10]). The AUC is 0.71, based on which we conclude that the test is fairly discriminatory.

[10] For example, a computerized prediction that churns out, at random, 'yes' in 15% of cases, and 'no' in 85% of cases. Such a test would meet the same false positive rate of ~15%. A 50-50 random test would predict 50% of cases correctly, but would also have a false positive rate of 50%.

Predicting ICU admission for ER patients is much more difficult. This is because while in planned admissions (e.g. for planned surgeries), ICU admission is part of the planning process, for ER patients the probability of ICU admission is much more random. We also see this in our analyses: at best, the test can predict ~38% of ICU admissions, if we accept a false positive rate of ~15%. The AUC for this curve is 0.56.

Overall, these results compare relatively well to other predictive models, although there are better models out there. For example, both Gagné et al.[11] and Loekito et al.[12] found AUCs slightly above 0.80 for their models. We would like to emphasize that our results are based solely on relatively naïve analysis of data points readily available in an EHR, without adding any medical knowledge or evidence-based decision support. We are convinced that adding more data points (e.g. by linking primary care data, pathology outcomes, radiology results, including historical data beyond one year, considering diagnoses based biases, etc.) and adding intelligence (e.g., have a physician input their own probability, and using that variable as one of the data points in the fingerprint model) could make the model more precise. These first results at the least encourage a follow-up on this methodology. But most importantly, the confirm that the wealth of experience data that is embedded within an EHR can be used to draw meaningful conclusions - even without applying any prior knowledge or expertise.

Outcome-based clinical decision support

Having established that the fingerprint model can draw meaningful conclusions on outcome based on the wealth of data that is collected in an EHR, we explored how the model might be used as a tool to support clinical decision making.

We hypothesized that the fingerprint model can be used to determine whether a particular clinical intervention has a meaningful impact on outcome for highly similar patients. If so, the model can either suggest treatment options based on fingerprint directly, or it can flag patients for whom recommendations from the model were not followed, e.g. for potential discussion with colleagues in grand rounds. Please refer to the section "Areas of application" for a more in-depth discussion of how the fingerprint can conceptually support clinical decision making.

[11] Gagné M et al. Performance of ICD-based injury severity measures used to predict in-hospital mortality and intensive care admission among traumatic brain-injured patients. J Trauma Acute Care Surg. 2016 Nov 30.

[12] Loekito E et al. Common laboratory tests predict imminent medical emergency team calls, intensive care unit admission or death in emergency department patients. Emerg Med Australas. 2013 Apr;25(2):132-9.

To test how this might work in practice, we considered a hypothetical situation where surgeons admitting elderly patients through the regular admission process (i.e., not through the emergency room) would like to evaluate whether administering oral antibiotics at time of admission contributes positively or negatively to length of stay.

Of course, such an analysis might be more interesting if we would consider specific kinds of antibiotics used as prophylaxis instead of a binary yes-or-no, and if we were to use more interesting outcome data, such as probability of infection during admission, overall survival, etc. Since such data were not available to us, we had to keep to a simpler model of which the outcomes themselves are perhaps not all that insightful. However, for our specific intent - to test how the fingerprint model might be put to use to support clinical decision making - the process of this particular analysis was more important than the outcome.

Analytically, the translation of this situation to the fingerprint model works as follows:

- We selected all test patients who were admitted by the surgical department, and who were admitted through the regular admission process. For the purpose of this analysis, the minimum fingerprint density was set at 0.001. A total of 94 patients were thus identified in the test set
- For each selected test patient, we identified a reference group with highly similar fingerprints. For the purpose of this analysis, the minimum similarity level was set at 0.96.
- We then separated the reference group for each test patient into 2 subgroups: 1 subgroup of reference patients who did receive antibiotics upon admission, and 1 subgroup who did not receive antibiotics upon admission.
- We compared the outcome (i.e., length of stay) for the reference group with antibiotics to that of the group *without* antibiotics, with the hypothesis that prophylactic antibiotics may, for certain patients, reduce the length of stay because they reduce the probability of post-operative infection. For the purposes of this proof of concept, we set an arbitrary threshold - if there is a difference in outcome of more than factor 1.5, the model provides a recommendation (in a more elaborate model, one might perform statistical testing between the groups instead of choosing an arbitrary factor). Otherwise, no recommendation is given.

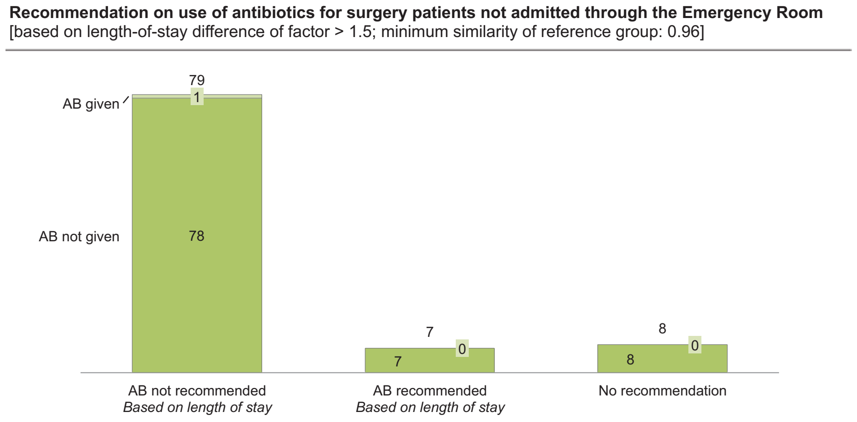
Figure 10 shows how often each type of recommendation is given vs. how often oral antibiotics were actually administered upon admission. Interestingly, oral antibiotics were almost never given in this test group of patients. For the one patient who did receive antibiotics, the model would have suggested not to do so. Upon further inspection, we found that for this patient the reason for admission was entered as 'unknown'. Review of activities performed on this patient suggest the patient was admitted for arterial occlusion issues with the lower extremity (this is also a likely explanation for why antibiotics were administered).

However, based on the outcome measure of length of stay, the model suggested considering antibiotics for 7 patients for whom they were not actually given. Upon further review of these 7 patients, we found that the reason for admission for 6 of them was surgery for rectal carcinoma (for 1 patient, it was lung cancer).

At this point, the model has flagged several patients with rectal carcinoma for whom the use of antibiotics might be considered, but where we know in hindsight that antibiotics were not given. The findings from this analysis, together with a more in-depth review of these patients' records and literature, might form an excellent basis for discussion at grand rounds. Indeed, a recent Cochrane systematic review concluded that the benefit of routine use of oral antibiotics is at present uncertain[13].

FIGURE 10:

Use of the fingerprint model for recommendation on use of antibiotics in surgical patients



The analytical steps demonstrated here can in principle be applied to any combination

[13] http://www.cochrane.org/CD001181/COLOCA_antibiotics-administered-patients-prior-colorectal-surgery

of clinical intervention and outcome measure. It can also easily be expanded to, say, compare the clinical decisions taken by different colleagues. Of course, in many situations, the insights derived from the model may not yet be meaningful without further detailed study and professional evaluation. But it is shown here that the model can at the very least potentially identify very specific subgroups of patients for whom it might be beneficial to discuss experiences among colleagues, perform a more detailed study, compare findings to evidence in literature, etc.

Summary of results

For the 2 fields of applications tested, length of stay and ICU admission, we have shown that the fingerprint model we have developed based on EHR data can already match or even surpass the best existing and published models. The fingerprint model, however, has other important practical benefits. Existing models typically use either a formula or a decision rule derived from multivariate regression analysis of a certain population. That means that applicability to other populations is uncertain. Our model uses 'live' data from the local EHR, so that it's predictive powers do not rely on certain pre-calculated parameters but on characteristics of the recent and evolving local population. The predictive power of the model evolves with the population and the care regimes. In that sense, it is a 'self-learning' model.

Furthermore, the fingerprint model can handle an unlimited amount of data points, while multivariate regression models are limited in the number of variables that can be incorporated.

We have, on purpose, not attempted to predict outcomes such as survival or quality of life. That is first because that data is not available in the dataset we have, but also because we wanted to first validate relatively non-controversial outcome measures to test the predictive capabilities. In the last analysis described above, we have shown how, when linked with outcome measures, one might use the fingerprint model to provide clinical decision support that leverages the wealth of available EHR data and applies it to the individual patient.

There have been many attempts at extracting predictive information from EHRs using big data techniques, and it was not our intent to prove that this model is better. However, as far as we know, there are no live working algorithms based on experience such as this one operating in EHRs today, and we do believe the fingerprint model is one of the first directly applicable models that has proven predictive power (for the two tested applications) and can readily and easily be implemented into any EHR system. Moreover, the model has the potential to achieve even greater accuracy with some minor improvements, which we will discuss more in-depth in the following chapter.

Improvements to the model

The previous chapters have described the background of this study, the design of the model and proof of its predictive capabilities. Before we expatiate how we can use this model in real-world applications, we will use this chapter to describe routes for improving the model, both from a conceptual as well as from a technical perspective.

Conceptual caveats to the model

Finding the synergies between doctors and computers – In the introduction we mentioned that doctors are largely reluctant to apply available information and big-data processing in their day-to-day work. This reluctance to admit automated data analysis into the consultation room can, in part, be traced to understandable concerns regarding the protection of individual patients' privacy. Another strong concern within the clinical community is that automated data analysis cannot and should not replace a balanced decision making process informed by years of training and experience. We agree that this is so. In fact, we believe data analysis techniques such as outlined in this paper should aim to supplement rather than substitute the existing, mature, highly developed and indeed very successful clinical decision making process. There are ways, in our view, that the human senses and brain function that cannot as yet, or perhaps even ever, be replicated by algorithms and computers.

Science is not democracy – The model essentially works based on one axiom: what happens most often is relevant and should be considered. The concept does not judge, does not provide any sort of verdict about which option is best; it merely provides the treatment options that are most frequent given the set of known patient characteristics. This in itself need not have any definite implication on treatment choices; well-informed and well-judging doctors may decide to omit information as often as they believe is in the best interest of their patients. However, there is a certain risk that prompting 'average values' could lead to less variation in treatment and the distribution degenerates to the mean. In other words, this effect may eradicate beneficial practice variation and hinder the search for better methods. There is a risk of enforcing majority rule rather than best and evolving practice.

For example: a patient with high cholesterol needs to be treated with a statin; a medication that lowers cholesterol. Imagine that there are three options: A, B and C. Their respective frequency is 60%, 30% and 10%. Our tool will give this information to the doctor, who may conclude that A is the best choice since it is most conventional. However, it is very well possible that C is a relative newcomer with far superior results. In an optimal scenario, the algorithm also incorporates predicted outcome. For example, in the statin case, for all options the expected quality of life, probability of adverse outcomes or survival rates in general could be reported. However, in any evolving field of application, and medical science is constantly innovative, there is always a time lag and past outcomes need to be judged in light of the evolving curve rather than the historical practices.

Technical improvement of the concept

Our goal was to develop a first feedback loop based on readily available 'experience' information. If the final vision is a modern full electric car, by analogy we now have a Flintstones car. But it's an artefact that can be used in everyday live, and is much faster than walking. The basis - the fingerprint algorithm - is the heart of our vision and can be extended and improved in various ways. Below, we outline several avenues for technical improvements of the first version we currently have.

First, we can introduce intelligent weighting to the different variables. Currently, we employ six different kinds of variables in the fingerprint: reason of admission, diagnosis, lab diagnostics, surgeries, other activities performed on patient and medication upon admission. The weights of these categories, as well as the values within these categories, are determined in a naïve way. That is, they are equal. This means for example, that the diagnosis 'diabetes' has the same weight as the result of an ALAT clinical chemistry test. The decay of weightings over time was also set at 1 year decreasing linearly, while the optimal weights may very well differ by variable. This naïve weighting can be made more intelligent by including current medical knowledge to tune and tweak the weight so that the suggestions of our model become more meaningful and insightful. Alternatively, we could allow users of the tool that is based on this model to indicate which predictors were relevant every time they use the tool. Machine learning algorithms can even be applied that extract which information leads to different medical decisions and continually improve the fingerprints.

Second, we can use readily available medical evidence in the fingerprints. For example, we know that methotrexate is used for inflammatory diseases such as rheumatoid arthritis, psoriasis and Crohn's disease. And it is not used for any other diseases. If we were to use our model for medication advice, the current iteration, though unlikely, might suggest the use of methotrexate for patients that have exactly similar surgeries, lab results, and reason of admission to the hospital as 25 other patients, but who do not have any of these diseases. Applying this common medical knowledge to the algorithm may exclude the suggestion of methotrexate for a patient without one of these diagnoses. Applying well known medical evidence to the fingerprints as rules or constraints will omit such issues.

Third, we can improve the significance of the suggestions of the algorithm by improving the cut-off points for the required density and similarity scores of the algorithm. For example, we can allow the algorithm to learn, by itself, what weights and cut-off points best suit the problem at hand. This could be done by linking the output of the algorithm with the actual decision of the doctor. With this link, the algorithm can observe what kind of information leads to which decisions. And optimize its weights and cut-off points given these new insights.

Areas of application

As we seek to make the world a better place, this chapter deals with perspectives about practical ways to bring our thinking forward and apply it in practice to truly have impact on clinical decision making.

Overall, we see three possible avenues of application for this model. These avenues and the concrete possibilities are outlined in figure 10:

- *Learning on the go:* diagnosis and treatment recommendations for healthcare professionals, based on clinical decisions made by colleagues for highly similar patients
- *The world is a clinical trial:* a new way of performing clinical research, that can deal with much greater variance in patient characteristics than typical

randomized controlled clinical trials and where results are applied in practice immediately

- *Management information:* for insurance companies and healthcare institutions

FIGURE 11:

Avenues of application for the fingerprint model

Clinical information	**Point-of-care insights**	• ALOS / ICU admission (detailed in previous section) • Medication optimization • Detailed surgery recommendations
	Macro-level insights	• Validation and optimization of existing evidence • Continuously adaptive clinical trials • Data mining for new insights
Non clinical information	**Management information** Insights for patients, governments and hospital administration	• Projected turnover and cost • Optimization of risk equalization schemes • Expected patient journey

Point-of-care insights

The fingerprint model readily lends itself to supporting a great variety of clinical decisions, both for treatment and diagnosis. Especially when combined with clinically relevant outcome measures, such as survival or quality of life, it could provide significant decision making support to physicians.

As you may recall from the introduction chapter, clinical decision making is a 3-dimensional problem. If there are *n* evidence data points, p patient characteristics and m experience data points, the problem is of order O*(m*n*p)*. The fingerprint model reduces this 3-dimensional problem to a 1-dimensional problem (the system only needs to compare 1 fingerprint to all other fingerprints), and therefore greatly reduces the complexity of inferring conclusions from all relevant data points.

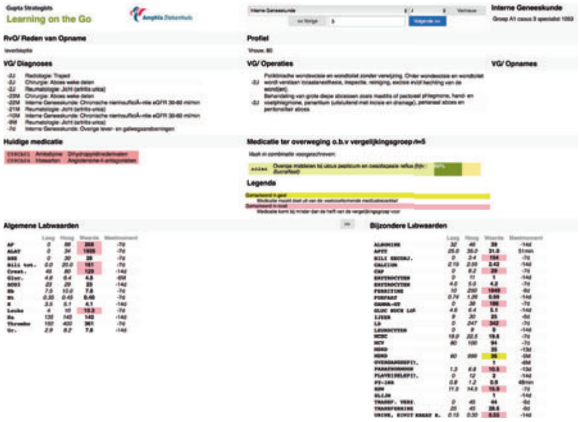
Thus, it is very well suited for clinical decisions where individual characteristics are likely to have a great influence on outcome, or where particular predictive variables are simply not yet known. We will here describe 2 contexts in which we think the model could be very useful, although there is of course an unlimited amount of other potential applications.

Medication advice for the elderly

Polypharmacy is the largest clinical problem in the elderly population. On average, 73% of people > 65 years of age use more than 3 separate drugs per day[14]. Elderly patients are also more susceptible to side effects and drug interactions. Admission to a hospital provides an excellent opportunity for a physician to review the medication list, but - apart from drug-drug interactions for which good databases exist - such a medication review is complicated and time consuming. Our model can be used to build an add-on to any EHR that provides instant medication advice to the physician based on what medication similar patients receive. We have actually built a proof-of-concept for such a tool, to show that it is 1) feasible and 2) quick enough to be usable. Figure 12 shows a Dutch version of this proof-of-concept tool. It provides an overview of the patient characteristics, such as reason for admission, lab values, etc. It also shows the medication on admission as entered by the admitting physician. The 'intelligence' added by the fingerprint model is revealed by the coloring of certain medications: a red highlight means the patients in the reference group typically do not receive that particular drug, while a yellow highlight means that the patient currently does not receive that drug but reference patients typically do. This is exactly how we envision an effective big-data driven clinical decision support system: a non-intrusive additional information point that leaves the physician in a better place to make the best possible decision for the patient.

FIGURE 12:

Proof-of-concept of a medication advice EHR add-on using the fingerprint model



Gupta Strategists

Learning on the Go

Interne Geneeskunde

RvO/ Reden van Opname

Profiel

VG/ Diagnoses

VG/ Operaties

VG/ Opnames

Huidige medicatie

Medicatie ter overweging o.b.v vergelijkingsgroep n=5

Legenda

Algemene Labwaarden

Bijzondere Labwaarden

[14] DEFENCE-II study, the Netherlands

Discovery analyses in detailed oncology databases

While the fingerprint model is very fast and is therefore particularly suited for on-the-fly analyses, it does not have to be used exclusively in such contexts. It can also deal very well with very large quantities of data points, and with situations where clinical knowledge does not yet exist: i.e., for what could be deemed "discovery analyses". It could, for example, be used to analyze large regional oncology databases such as the Dutch DICA[15] databases. Fingerprints could incorporate detailed tumor characteristics (such as Breslow thickness for melanoma), and the model could be used to identify fingerprints of patients that, for example, respond particularly well to specific kinds of chemotherapy. Researchers could then analyze this identified subgroup of patients to identify clinically relevant hypotheses for further study. As producers of medical technologies (devices and consumables) increasingly focus on value based healthcare, we see a role for them as a catalyst or champion in taking our model forward.

There are of course many more conceivable examples where the fingerprint model could be exceedingly useful.

Macro level insights

The previous paragraph is about softly 'nudging' or 'guiding' decision making for doctors. It acts on a micro level: it aids decision making on a patient basis at the point of care. However, we can also take a more macro point of view. By not focusing on one single decision but a string of multiple decisions over time - the algorithm gets a whole new dimension: it allows us to see patterns over groups of similar patients and over time. In this way, the model may become an input tool for faculty discussion meetings: it can for instance signal variation between young/experienced clinicians, and between various educational backgrounds (different countries, different universities, etc.). And it can be highly valuable to preselect areas where practitioners can learn from each other by signaling between doctors. Three manifestations of this application are outlined below.

Validation and optimization of existing evidence

Most of the clinical evidence is based on so called randomized controlled trials (RCT). RCT method is important to ensure validity of claims but by the same design it limits the applicability of the claim. The problem lies in the 'controlled' part of RCT: criteria

[15] Dutch Institute for Clinical Auditing

for admission to an RCT are very strict in order to allow for statistical inferences about outcomes. This has the advantage that with quite low numbers of patients, researchers can often already make claims about, for example, the effectiveness of a new therapy. The drawback is that these criteria make the reference group so narrow that it is almost impossible to have a 'real life' patient match exactly all of the criteria of an RCT. Therefore, it opens the possibility that observed effects in RCT are not fully replicated in the field.

Take, for instance, the study that compares cream therapy of a basal cell carcinoma against the current best practice of photodynamic therapy and against surgical excision (Roozeboom[16], 2016). The study concludes that in many cases, cream therapy is preferable to photodynamic therapy.
However, the criteria are strict: only patients above age 60 are included in one of the subgroups; only carcinomas that are superficial are included; only patients from the south of The Netherlands were included (mostly Caucasian); patients were only included from 2008 to 2010; and all tumors not in the head / neck area were being treated as one group. Furthermore, there were very specific choices about the frequency and intensity of the treatments, and the way diagnoses were determined was also very specific.

These types of restrictions are quite common and necessary to design a well-defined trial. But the implication for practice is very discouraging: almost no patients meet the criteria used in the RCT exactly. What if there's a patient who has an almost-but-not-quite superficial basal cell carcinoma on the shoulder, is 59 years old, has Hindustani background and also suffers from psoriasis? While the RCT is conclusive about the highly particular patient group it studies, it tells us nothing directly about this individual patient. Designing and running RCTs for all kinds of segments is prohibitively expensive, even the current restricted trial regime is hardly cost sustainable. Finding ways of relaxing the assumptions and criteria of RCTs would be an interesting proposition.

The fingerprint model with similarity scores can help improve study designs to do exactly this: make study outcomes more generally applicable, albeit at the small extra expense of less strict inference. This disadvantage can be solved by increasing the number of patients, and possibly also by advanced statistical procedures such as bootstrapping. The way a similarity score can substitute conventional criteria in RCT is quite simple from a design perspective.

[16] Roozeboom et al., Three-Year Follow-Up Results of Photodynamic Therapy vs. Imiquimod vs. Fluorouracil for Treatment of Superficial Basal Cell Carcinoma: A Single-Blind, Noninferiority, Randomized Controlled Trial. J Invest Dermatol. 2016 Aug; 136(8): 1568-1574

First, one determines exclusion criteria (for example, some diagnoses or age groups may be excluded). Second, one determines the 'model patient' and cut-off value for similarity score with the model patient (for example, >0.95 similarity score is required). Third, one computes similarity scores for patients to be included in the study. When interpreting the study outcomes, an additional robustness check might be performed for patients that are outlier in one or more dimensions.

Continuously adaptive clinical trials

There is another way to learn from the outcomes and influences of the algorithm. In healthcare, substantial time series and panel data are generated and captured in various databases. These data contain information about patient characteristics, chosen treatments and quality outcomes. These extensive data points enable researchers to adopt new information earlier and inflict less damage to test subjects that unnecessarily receive inferior treatments.

This information can be used to alter the way clinical trials are executed entirely. As mentioned before, the issue with RCTs is that they are quite inert and inflexible: there is no learning in the process. One example is the story of ECMO (Extracorporeal Membrane Oxygenation), a technique to provide oxygen to the body for persons whose heart and lungs are incapable to do so. When this technique was new, a series of RCT's were performed in different locations and in slightly varying populations. In total 39 newborns deceased in five different experiments and studies, because they did not receive ECMO. The success rate for ECMO in those studies was near 100%, also for the earlier ones.

There is a branch of clinical trials that allows for fast learning. It relies on the following premise: as long as the distribution of potential outcomes of a new treatment includes outcomes that are better than the outcomes of the current standard of care, the new treatment should be chosen. If a treatment has never been administered, there is for instance a 50% chance that the new treatment is better, and a 50% chance that it is worse than the current treatment. Of course, a 50% chance that it is better is quite significant, so the premise implies that we should try the new treatment. Once we have tried the treatment, we have more information on the actual distribution of outcomes, but the premise still holds: choose the new treatment as long as there is a significant chance that it might be better.

Eventually, either the distribution of outcomes is clearly in favor of the new treatment, and we should reject the old standard of care, or the new treatment is clearly inferior and we should never administer it again.

For example: if 5 newborns were treated with ECMO and 5 without. Of those with ECMO, 4 survived. Of those without ECMO, 1 survived. The numbers aren't sufficient for statistically robust results. However, there is some information in the results that researchers can leverage on - the premise would clearly imply that the next patient *should* receive ECMO because it is significantly better than the standard of care. This approach, rather than the repeated RCT's, could have saved many lives.

This type of trial, which can be grouped under the evolving concept of the "adaptive trial"[17] and which me may therefore dub "continuously adaptive trial", essentially converts everyday clinical practice to a clinical trial.

It is particularly suited for measuring the effect of minor clinical decisions that are normally not subjected to clinical trials, such as alterations of the salt level in hospital catering, or the choice between several different statins for treating high cholesterol.

Where does the fingerprint model come into action? The fingerprint model can be used at any time during such continuously adaptive trials to observe and analyze effects. Rather than asking "what is the distribution of outcomes for this treatment?" it allows us to ask "what is the distribution of outcomes for this treatment in highly similar patients?". For any outcome measure, there may be variables that greatly influence success rate but that are as of yet unknown - whether it be a specific age group, anatomical differences of the heart or comorbidity, etc.

By enabling fast, on-the-fly analyses of the entire EHR database and inferring conclusions on outcome distributions for highly similar patients, our model allows hospitals to continuously improve the quality of care for all patients. Essentially, it allows for the conversion of "continuously adaptive trials" into "continuously *controlled* adaptive trials".

Data mining for new patterns

Analogous to using similarity scores in new study designs, the algorithm also lends itself to retrospective data analysis. For example, once a large data set - such as an EHR - is obtained, researchers can compute similarity scores for all pairs of patients or for all patients against a 'model patient' they have in mind.

[17] For an excellent background on this topic, please refer to: Bhatt et al., Adaptive Designs for Clinical Trials. N Engl J Med. 2016 Jul; 375(1): 65-74

Then, researchers can test their hypotheses in this data set. For example, testing treatment A against treatment B for a specific diagnosis. Of course, pure randomness is not possible by design.

By only comparing patients with very high similarity scores, a very large degree of randomness could be engineered. For example, consider taking a patient from group A (who received treatment A), that patient can be matched with the best matching patient from group B by calculating and ranking all similarity scores from group B with the specific group A patient. This process can be repeated until no more matches can be made above a predetermined threshold similarity score.

Regression analysis would be a natural alternative that comes to mind. This need not be a substitute for similarity scores. On the contrary, similarity scores may complement regression analysis. For example, by excluding highly dissimilar patients before starting the regression analysis.

Management information

As well as providing clinicians and researchers with valuable new insights, the model can also provide managers, administrative staff and even patients with previously unavailable insights. Below, we outline three avenues our model can generate new perceptions.

Improving risk equalization schemes used by healthcare insurance companies in NL

In The Netherlands, the health insurance system is strictly regulated. One important feature is that it is obligatory for inhabitants. Another feature is that insurance companies are obliged to accept any new participant. Of course, this results in different risk profiles of insurance providers, since they are active in different parts of the country and have differing market shares for age strata and social status strata. The government has a system in place to correct for these differences in risk. This scheme works quite well for population as a whole. But for specific sub segments, it works but moderately, and for individual patients, it has hardly any explanatory power at all. [18]

[18] See, for example: WOR 748 - Onderzoek Risicoverevening 2016: Overall Toets. Table 2.35 (page 69). Instituut Beleid en Management Gezondheidszorg.

The fingerprint model can help predict individual health care costs in the following novel way. First, it can match the insured for whom you want to predict the cost of care with all other highly similar insured (based on the similarity score) - without requiring any up front knowledge of what particular characteristics matter the most. Then, it can take the retrospective median or average of actual cost of care of the generated reference set. Of course, this requires a fairly large dataset. However, complete data sets of the entire Dutch population that would allow for such analyses have been meticulously collated and are readily available.

Patient in charge

Patients are becoming increasingly informed about the goods and services they consume and with this shift the patients' expectations grow further. For example, airlines have mobile phone apps that show the products, times and delays. They even offer optional customized extras such as extra leg room or a meal to order. Equivalently, Amazon has an app that provides all past and future purchases, provides a wish-list with its own suggestions, and keeps track of delivery information. Consumer interaction possibilities and uptake is continuously on the rise. Hospitals and healthcare in contrast seem to be caught in a pre-historic time warp. The records are mostly kept in paper form, appointments are made by telephone and confirmed by traditional mail, if at all. And patients are hardly informed about what procedures they can expect, at which location and at which time.

Our model obviously can't change that hospitals are relatively slow in their transformation to the digital age.

However, it can help hospitals to give patients an 'average' prediction about what care activities they can expect and what the involved costs will be.

This can be done by creating an 'expected care path'. A care path in this case is a set of care activities that a given patient receives, for example: the different lab requests, MRIs, surgery and a stay in the hospital. Based on the set of historical care paths, the model can calculate the expected care path by taking the modal care path from the set of patients who have the highest similarity score.

Predicting production volume and revenue for hospitals and insurance companies

Hospitals and their payors - the insurance companies - have prediction models for their cost and turnover. Due to the nature of the payment system in The Netherlands, these models tend to have quite a big time lag. Hospitals in the Netherlands are paid for 'packages' of care that frequently go together. The definite 'package' for which a hospital will be reimbursed by the insurer will only be known after the completion date for the entire package, which can be many months after the first diagnosis. For example, when a patient visits the dermatology outpatient clinic with certain symptoms, there are very often many different possible packages as outcomes. And even as the treatment advances, there is much uncertainty of the eventual package. An extra MRI can change the package to a more expensive one.

Of course, there are prediction models to handle such uncertainties. However, they infer production based on historical patterns of the consumed aggregate care, and not on a per patient basis. Our model can predict expected care (following the analogy of 'patient in charge') and hence expected cost for an individual patient.

[19] Berner, Clinical Decision Support Systems: State of the Art. Jun 2009, Agency for Healthcare Research and Quality (AHRQ) Publication No. 09-0069-EF

Appendix – background and overview of clinical decision support

"Clinical decision support (CDS) systems provide clinicians, staff, patients, and other individuals with knowledge and person-specific information, intelligently filtered and presented at appropriate times, to enhance health and health care"[19]

While this definition is concise and captures the purpose of a CDS, it doesn't mention how a CDS works under the bonnet. Clinical decision support is the brains behind any system that healthcare professionals use in care delivery. Clinical decision systems are an evolution on the development of electronic health records. EHRs, e-prescribing systems, computerized physician order entry, and medication reconciliation systems all are strengthened by some form of clinical decision support. CDS can help physicians reach proper diagnoses, ask the right questions, and perform appropriate tests on the front end of the decision-making process - preventing errors of omission - as well as stop errors of commission on the back end, during treatment and procedures.

However, in reality a CDS is a multifaceted animal that takes on many different shapes. Hence, this paper provides a more practical approach to explaining CDS. Basically, any analysis can be structured in terms of its input, its computations and its output. This holds true for CDS as well, so we'll explain the mechanisms of a CDS along these lines. Then, the most important CDS are compared along these elements.

Overview of elements that differentiate CDS

Input - CDS use different kinds of input. First of all, they use patient variables: demographic properties such as age and gender, clinical variables such as complaints and pain and biomarkers such as potassium or X-ray results. Secondly, they use variables on care activities performed on the patient, such as the amount of OPD visits, medication usage or type of surgery performed. Thirdly[20], the CDS may use outcome variables, such as pain score. Lastly, it may use the use clinical guidelines on diagnosis and treatment, from national guidelines to hospital specific guidelines.

Calculation - The way the CDS use their input to calculate varies tremendously. They all rely on a combination of algorithms that attempt to identify the best possible diagnosis or treatment. The simplest form is a straightforward correlation analysis. The most complex variants entail a combination of optimization algorithms (such as k-means and

[20] The information in this chapter is partly based on several articles from the InformationWeek website

J86 step-wise regression). The table below provides some examples of those algorithms to give a better idea.

Output - There are many distinguishing elements that typify the output of a CDS. The most important one is the point in the care chain that the CDS applies to. There are two categories. On the one hand, there are CDS that help in diagnosing a patient. As a doctor enters more parameters, the CDS calculates the most likely disease and asks relevant additional variables that increase the precision of the prediction.

On the other hand, there are CDS that help determining the best treatment for the patient. For example: is it - given all the specific data points - better to treat a patient with medication or to perform surgery?

Apart from the point in the care path (diagnosis or treatment), there is a number of other, though less discriminating, properties: the type of information it displays, the level of detail that it gives, the point in the decision making process it provides information and the person to whom the CDS reports.

The type of information - does the CDS give information about variation between doctor's choices? Does it also give information about the relation between the choice and health outcomes? Does the CDS specify exactly which route a doctor could follow, or does it provide high level information? Is the information descriptive or normative, ie. does the CDS steer in a certain direction?

The user of the program - does the CDS report to employees involved in patient care such as doctors, nurses or other healthcare workers? Or does the CDS target managers or directors of healthcare institutions? Of course a combination of these is also possible: a CDS could guide doctors at the point of care and also provide overall business intelligence to managers.

The point in the decision making process - does the CDS provide information before or after the decision takes place. There are CDS that give live feedback, ie. they provide information that helps make decisions at the point of care. On the other hand, there are CDS that enter the process ex post. One would think that this is less valuable than on-the-spot feedback. However, ex post feedback allows for analysis of decision making and arranges for additional insights in the decision making process.

Name of CDSS	Developer	Diagnosis / Treatment	Point of care	Input	Algorithm	Output
Archimedes	Kaiser Permanente	Treatment	Yes	Clinical history (comorbidity, lab values, patient properties etc)	Automated regression analysis	Treatment options including expected outcome
Auminence	Cambridge, now owned by HP	Diagnosis	No	Patient history, presenting sympto-ms, physician knowledge	Analysis of patterns and correlations in data	Checklist to arrive at the correct diagnosis, icluding statistical probabilities
SmartPath	Diagnosis one	Treatment	Yes	All available patient data	Scans data to identify gaps in care	Gives ideas on best treatment as the doctor enters more information
DXplain	Mass General Hospital	Diagnosis	Yes	Symp-toms, medical evidence, physician observatios, test results	Scans appropriate medical evidence	Possible diagnoses
Dr. Waton - Watson-paths	IBM	Diagnosis	No, meant for education	Clinical evidence	Text recognition, analysis of patterns and correlations	Possible diagnosis incl. statistical probabilities
Advisor	PKC	Treatment	Yes	All available patient data	Matches patient profiles to relevant medical evidence	Treatment options based on best fit between literature and specific patient

For further information on this study, please contact:

Kees Isendoorn
kees.isendoorn@gupta-strategists.nl

Daan Livestro
daan.livestro@gupta-strategists.nl

Gupta Strategists

www.gupta-strategists.nl