Towards Consent-Based Lifelogging in Sport Analytic

Håvard Johansen[1], Cathal Gurrin[2], and Dag Johansen[1]

[1] UIT The Arctic University of Norway
[2] Dublin City University, Ireland

Abstract. Lifelogging is becoming widely deployed outside the scope of solipsistic self quantification. In elite sport, the ability to utilize these digital footprints of athletes for sport analytic has already become a game changer. This raises privacy concerns regarding both the individual lifelogger and the bystanders inadvertently captured by increasingly ubiquitous sensing devices. This paper describes a lifelogging model for consented use of personal data for sport analytic. The proposed model is a stepping stone towards understanding how privacy-preserving lifelogging frameworks and run-time systems can be constructed.

1 Introduction

Wearable and ambient lifelogging technologies that individuals intentionally use to capture aspects of their own activities, promise life enriching benefits like heightened self-awareness, personalized health-care applications, and new ways of learning. This might lead to longer and more active lifespans, increased productivity in the workplace, increased independence, or increased mobility for people suffering from various memory and cognitive impairments. Lifelogging also fosters new forms of social interaction and sharing [4]. We are already seeing applications in triggering recall of recent memories. This is an application of lifelogging where the detailed lifelog acts as a memory prosthesis, thereby providing support for people with Alzheimer's or other forms of dementia [2,10,16]. Extending this concept from the care-giving domain to every-day life, there is potential for lifelogging to provide memory support to fallible human memory [6], or for logging of activities in molecular medicine [15].

The *quantified-self* movement [18] is perhaps a first mass-scale instantiation of these technologies and has lead to the emergence of pervasive lifelogging as a mainstream activity where individuals use automated digital sensors [8] to capture and permanently store a comprehensive unified digital archive with data related to their lives. An overview of the different categories of lifelogging tools that have been employed can be found in the work of Machajdik et al. [17]

In this paper, we are concerned with the specific case of using lifelogging as a tool for improving coaching and avoiding injuries in professional sport. Collecting, storing, analyzing, and correlating large volumes of personal data, known as *big-data analytic*, from teams of athletes, is important for the emerging next

generation sports analytic systems. Such systems enable coaches and medical staff to find useful performance and health indicators that might not be visible from studying single records alone [20]. This has the potential to detect individual health and performance problems at an early stage so that coaching and exercise programs can be individualized, which is clearly highly beneficial for both the athletes and the elite sports clubs.

In collaboration with Tromsø Idrettslag (TIL), a Norwegian professional soccer club, we have developed several lifelogging tools and systems specifically targeting this particular domain [9, 12, 13]. As such tools are becoming more common, comprehensive, and continuous, new privacy concerns are coming to the fore and rise important ethical and legal problems [1].

Some data, such as ambiently recorded images and audio, will naturally pose more concern than others. Indeed, privacy is an inherently fuzzy concept [19] and has had many meanings, definitions, and expectations that differ across jurisdictions, areas, and over time. Also, privacy is affected by political, social, and economic changes and by technological developments; and include psychological, social, and political aspects.

While ambiguities inevitably arise, this paper derives a simple model for reasoning about such privacy using the principles of attribution and access to captured data. This model is derived directly from our real-world experience from applying lifelogging tools for both personal and professional use in elite soccer. Although our model does not cover all privacy related issues, it forms a useful initial framework in which more complex models can be explored.

2 Lifelogging in Sports: A Use Case

Elite sport is a fiercely competitive domain where technology is currently being widely adopted as a game changer. Lifelogging has in particular surfaced as a potent tool for athlete quantification, reshaping how sports are played and how athletes are being developed.

In our case, we have developed and deployed several prototype systems for lifelogging in elite soccer clubs in Norway and for the Norwegian national soccer team. Figure 1 illustrates specific prototype deployments related to the elite soccer club TIL, enumerating the different lifelogging components. As seen in the figure, each individual athlete carries FitBit Flex armbands 24/7 (1), and he also reports perceived wellness and fitness data on a daily basis. This includes parameters like, for instance, perceived fatigue, sleep quality, and muscle soreness, and this must be manually submitted every morning before 9 AM through a smartphone app (2).

Similar reporting is also done post-practice, where perceived personal training load (RPE, perceived exertion on a category 10 scale) is submitted (3). A central server based on the Ohmage platform collects and stores this data, and results of statistical queries can be graphically depicted for coaches and physicians (4). Additionally, medical staff and physicians collect test results periodically for individual players. This includes pure medical stress tests (for instance measuring

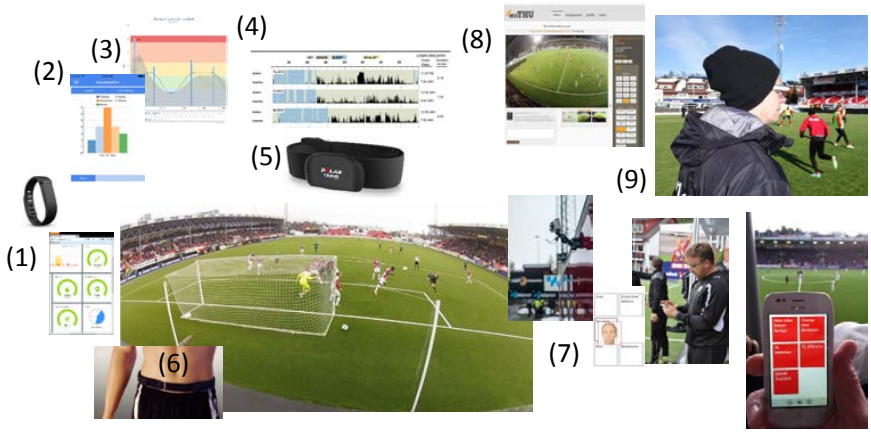

Fig. 1. Lifelogging equipment used for athlete quantification in TIL

lactate threshold predicting athletic endurance performance) and use of jumping boards measuring how high an athlete jumps.

Supplementing devices and services can be used during a practice session. One is Polar pulse belts strapped on each player (5), a wearable monitoring device wirelessly connected to a tablet carried around on the field by the club physician. This way, the individual athletes' pulse is constantly monitored and can be used for interventions (for immediate personalized lower or higher load adaption). Another system used is ZXY Sport Tracking (6), a radio based body sensor network system computing and storing positions and physical data of players on the soccer field with a resolution of up to 20 samples per second.

Bagadus [9] is a novel video processing and real-time smartphone-based notation system we have developed (7). This provides video footage of specific events tagged at run-time, which can be used for feedback purposes instantly, in the intermission break in the locker room, or for post-game analysis where players get involved through a social network service (8).

As illustrated in this use case, lifelog data originates from a wide-range of sources ranging from wearable *inward looking* sensors, like those positioned on the body to monitor heart-rate and lactate, to the new generation of *outward looking* wearable devices that has matured and come to market. These devices incorporate wearable cameras (among other sensors) and can capture detailed photo and video logs of a person's daily activities in an automated manner. Devices like the Narrative Clip wearable camera or Google Glass, enable us to record video of every waking or sleeping moment of the athletes (9). The usage of such recording devices is of great interest to sport analytic because it can be used to capture factors, like food intake and environmental effects, that might influence the athlete's restitution outside of the training and game arena.

The use of such technologies is not particularly problematic in simple usage scenarios, like when a Google Glass is being worn by a soccer coach to capture

activities during training session (9). However, once worn outside of the controlled sphere of the station, such devices will inevitably capture highly personal and sensitive aspects of the lifelogger's activities, like what you are reading or toilet visits, as illustrated in scenes A & B of Figure 2. Even more problematic, is the capturing of the images and activities of other individuals in the form of colleagues, family members, friends, as illustrated in C & D of Figure 2.

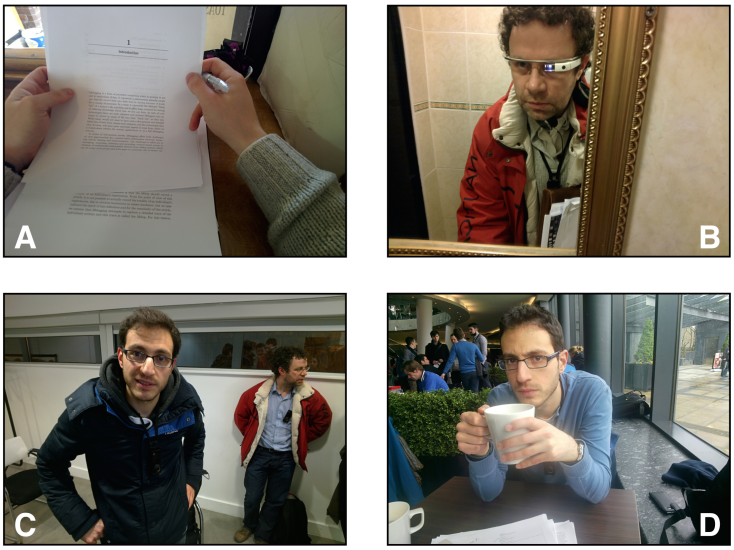

Fig. 2. Examples of potentially private data (A & B) and bystanders (C & D) in a Google Glass captured lifelog from 2014

The inclusion of such data in sport-analytic performed by coaches and staff in the sport clubs therefore might infringe upon both the privacy rights of the athlete and the privacy rights of inadvertently captured bystanders.

3 Consent in Sport Analytic

We have a principled approach to *privacy-by-design* and all personal data in analytic projects should have a strong notion of being voluntarily contributed. As the particular type of lifelogging described in our use case will often capture highly sensitive and personal data, we adapt to our purpose a fundamental definition on privacy based on the one by Dodge and Kitchin [5]. We considers the privacy for athletes to be based on

1. the right to choose the composition and the usage of your lifelog data, and
2. the right to choose what happens to your representation in the lifelogs of others.

A key problem for professional athlete quantification is that data capture is not initialized and controlled by the individual himself, but rather managed by a large number of people in the form of coaches, physicians, and support staff that typically surrounds sport clubs. These people typically manage the sensor systems that collect data on the individual athletes and specify which analytic functions to run. The rights for a sports club to captured and analyze data must therefore be attested formally as an explicit stated *informed consent*. It is required that such consents are an individual's autonomous authorization of an intervention or participation in some project where private information might be disclosed and used [3].

Informed consent from subjects of study, is already a well established concept in the field of medicine. Physicians or medical researchers requiring personal data have an obligation to tell the subject about the procedure of the participation, the potential risks, and benefits to the subject. The subject must not be deceived or coerced, which implies that each subject must adequately comprehend the consent they are asked to give. Finally, the subject must intentionally sign the consent form.

Traditional consents in the form of passive paper documents, stated and signed at the time data is first collected, does however not capture well the dynamic protection and consensual agreement needed for long term usage of lifelogging data in sport analytic. Emerging data mining algorithms that can discover new sensitive personal traits from existing data or organizational changes in sport clubs, might change the mind of athletes in what they to provide and for what purpose. Using paper-based consent forms, the fine-grained consent management required to support such fine-grained control of personal data is a daunting task.

4 Modeling Lifelogging for Privacy

As highlighted in our use case in Section 2, we are deploying sensors and devices to systematically capture, in digital form, a finite set of personal attributes for each athlete in the sport club. To express this, we have adopted a simple data model that captures data from each lifelogging device d as sequence of measurement samples

$$s_i = [d, t_i, \delta_i, \boldsymbol{v}_i] \ , \quad i = 1, 2, \ldots$$

Here t is a unique monotonically increasing number for source d denoting the time when the sample is recorded, and $d = (type, device)$ identifies data class contained in the record, like "pulse" or "step" in combination with the identity of the device that generated the data, like `zxy.belt.13` or `RunKeeper.app.78a9fac2`. By explicitly stating device names, multiple devices that provide similar types of data can be supported. For instance, both the ZXY Sports Tracking (ZXY) system and the Polar Belts provide pulse data. The value vector $\boldsymbol{v} = [v_1, \ldots, v_l]$ denotes l source specific measurement points for the sample, and may contain

arbitrary data like integers, text strings, or even large binary objects like images and sounds.

In addition to the time-stamp t, each sample s includes a time offset δ that indicates the time-span $[t, t + \delta]$ for when s is valid. For instance, positional data from modern 10 Hz GPS device, $\delta = 0.1$. A value $\delta = 0$ indicates the end of a sequence of samples, which is used to distinguish time-spans with no samples from time-spans between two valid samples. We denote the superset of all recorded samples as $\mathcal{L}$, the aggregate lifelog of our sport club.

Recorded data samples in $\mathcal{L}$ are stored in a digital archive that spans multiple personal computers, proprietary systems, and Internet services. Moreover, each computer system storing samples may do so for multiple individuals. For instance, in the ZXY sport-tracking system, a single database table holds entries for many athletes. This is also true for most modern solipsistic lifeloggers, as many popular sensor devices, like the FitBit Flex or the Narrative Clip, are hardwired to upload data to the vendors' shared data store. As such, the lifelogs of the principals p, $\mathcal{L}_p$ cannot be defined solely by the devices and systems storing his data, but must instead be defined relative to how data is related to the individual athlete.

4.1 Attribution of Data

To find a method for modeling how individuals' lifelog can be related to the data captured in $\mathcal{L}$, we turn to the notion of data attribution. Let $I(S) = P$ be some function that maps a set of data samples S to some set of principals P. This explicit mapping function I models the fact that most data types do not directly encode information that identify individuals. Clearly, a singular heart-rate sample like (6:10 pm,130 bpm) cannot by itself be attributed to a specific individual. The extra meta-data required for correct attribution, or the means to obtain it, must therefore instead be encoded as of this assignment function. Moreover, we say that I is *p-correct* if principal $p \in P$ would agree to the mapping upon manual inspection.

With this we define attribution, in the following manner:

Definition 1. *A data sample s is attributable to principal p if and only if there exists some p-correct mapping function I such that $s \in S$ and $p \in I(S)$.*

This gives us the following definition of a lifelog

Definition 2. *The lifelog $\mathcal{L}_p$ of principal p is the subset of all samples in $\mathcal{L}$ attributable to p.*

Correct attribution of data samples to individual principals is axiomatic to acquiring informed consent and therefore to privacy. To see this, consider a data sample s correctly belonging to p_1 but incorrectly attributed to p_2. This sample will not only damages integrity of $\mathcal{L}_{p_2}$, but also the confidentiality of p_1. Although correct attribution is a requirement for confidentiality it is not sufficient. This because we allow I to return more than one principal and do not require s to only

be attributable to p. To see this, consider our use in Section 2 with stationary cameras on the soccer pitch. Each captured image might be attributable to multiple soccer players and therefore be included in multiple lifelogs, potentially leaking sensitive information. We can formulate this property for consent and privacy in lifelogging as the following

Property 1. $\forall s \in \mathcal{L}_{p_1}$ where there exists some attribution function such that $\{p_2, p_1\} \in I(s)$ and $p_2 \neq p_1$, then p_2 have consented to the storage of s.

4.2 Data Access

Another key requirement of granting informed consent is that the recorded athletes can know about what is captured and what the data is being used for. Indeed, any data sample s that is attributable to p, but that p does not know about, violates privacy because the p cannot consent to its acquisition and use.

A key problem granting such individual detailed insight is that many athlete quantification systems do not provide direct access to low-level sensor data. This is also true for many consumer devices used for inward looking lifelogging where the sensors, like the Fitbit Flex armband, are often hardwired to upload data directly to some online backend server owned by the hardware vendors. These hardware vendor typically only present derived values.

The privacy requirement of informed consent fortunately does not require athletes to have access to the raw sensor data. Indeed, this might even be counter productive as low-level signaling data rarely gives insight in the captured data. The purpose of these systems is to improve athlete performance and prevent injuries and athletes may gain attributable output derived from their data in the form of individualized training and exercise programs, or high-level performance reports [14].

We therefore include in our model of data access, the relaxed constraint of only having the ability to read the output derived through some analytic function T on attributable data samples. We then obtain the following:

Definition 3. *Given a set of data samples S such that $\forall s \in S, P \in I(s)$. Then S is accessible to P if there exists some transformation $S' = T(S)$ such that $\exists s' \in S'$ where $P \in I(s')$ and s' is readable by P.*

In medical research, similar usage characteristics can be found in epidemiological studies on larger population cohorts, like the Tromsø Study [11]. Here, the usage of body sensors and other lifelogging tools is gaining in popularity, enabling more accurate longitudinal data to be collected. Such studies have previously relied on anonymization and de-identification techniques, like crowd blending [7], to preserve the privacy of the subjects. However, as personalized intervention technologies are introduced based on real-time backend analytic over collected data sets, the ability to attribute high-level output back to the individual data donors becomes a key function.

Having the definitions of access we can formulate the following axiomatic property for consent and privacy in lifelogging:

Property 2. $\forall s \in \mathcal{L}_p$, s must be accessible to p.

4.3 Bystanders

Athletes wearing outward looking sensors and engaging in normal every-day interaction with other individuals, like when attending work or socializing with friends, will inadvertently lead to situations where an attributable principal $p \in I(s)$ is not granted access to s—he might not even be aware that he is being captured. This is a highly undesired situation as without access, p cannot consent to being captured and stored and thus storing and using s might damage his privacy. We denote such principals as bystanders, defined in the following manner:

Definition 4 (Bystander). *Let s be a data sample from some lifelogging device worn by principal p_i. Any principal $p_j \neq p_i$ captured, either inadvertently or on purpose, by s such that s is attributable to p_j but not accessible to p_j, is said to be a bystander to s. If s is accessible and attributable to p_i then we also say that p_j is a bystander to the lifelog L_{p_i} of p_i.*

Depending on the lifestyle and sensor devices in use, bystanders could be predominantly strangers or they could be work colleagues, family members, or friends. It is expected that bystanders will be present in a significant fraction of the data samples captured using outward looking devices. Having this definitions we can state another axiomatic property for consent and privacy in lifelogging as:

Property 3. $\forall s \in \mathcal{L}_p$, s must not contain bystanders.

5 Discussions

Publication of lifelog data with bystanders is perhaps the most obvious privacy violation. All stages of the data life cycle have their concerns. For instance, in jurisdictions such as Japan, simply sampling pixels from a worn camera without consent of a bystander can be considered a breach of privacy.

Another interesting aspect of lifelogging devices and also a criticism that has been leveled at it, is that it is primarily a Write Once Read Never (WORN) technology. The view proposed by Bell and Gemmell in Total Recall [4] is that one never knows when some piece of data could be very valuable. This view is certainly also prevalent for data analysts in the sports domain, but is certainly not without its issues. In Norway, for instance, data retention laws that restrict such desultory storage of attributable data are already in place for non-personal use. In lifelogging frameworks that only evaluate data-access policies during capture, such restriction are important for preserving privacy as a lack of an explicit plan on its usage undermines the ability of bystanders to give informed consent.

We must also consider that lifelogging is typically carried out ambient or passively without the lifelogger having to initiate recording. This causes the problems of non-curation or non-filtering, in which the individual lifelogger may not even be aware of all the data being recorded, or the implications of keeping it. Also, due to sheer data volume, there is no practical opportunity for the lifelogger

to manually curate his data post-capture to remove or resolve any potential privacy concerns. For instance, the dominant visual lifelogging sensor used thus far is the wearable camera that passively can capture upwards of 4,000 images per day. We therefore propose that a lifelog can store indefinitely the complete representation of the activities of the lifelogger. Still if there at some point in time emerge an identifiable tort as a result of publishing the representation of an individual, then that individual should have the right to remedy. Hence, privacy preserving lifelogging assumes the presence of some mechanism that can postpone the acquisition of consent from the capture stage and until the later stage when derived data is accessed.

6 Conclusions

Lifelogging is a phenomenon whereby people can digitally record personal data in varying amounts of detail, for a variety of purposes. In a sense, a lifelog represents a comprehensive archive of a human's life activities and offers the potential to mine or infer knowledge on those activities using a multitude of software and sensors.

As we move away from naive implementation of lifelogging frameworks, where individuals solipsistic gather data into private life archives, to shared environments where collected data is used for purposes transcending the individual, like improving the performance of a soccer team, putting in place an appropriate privacy-preserving framework becomes an imperative. Towards this end, this paper have defined attribution as a key element of data in lifelogs, and argued that access to attributable data is a fundamental property to privacy. We also identify the presence of bystanders as a key problem, which can only be addressed by either deleting conflicting data samples or acquiring consent.

Acknowledgments. This work has been performed in the context of the iAD center for Research-based Innovation project number 174867 and FRINATEK project number 231687, funded by the Norwegian Research Council.

References

1. A. Allen. *Dredging-up the Past: Lifelogging, Memory and Surveillance*. Scholarship at Penn Law. University of Pennsylvania, Law School, 2007.
2. R. M. Baecker, E. Marziali, S. Chatland, K. Easley, M. Crete, and M. Yeung. Multimedia biographies for individuals with alzheimer's disease and their families. In *2nd International Conference on Technology and Aging*, 2007.
3. T. L. Beauchamp and J. F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, 2009.
4. G. Bell and J. Gemmell. *Total Recall: How the E-Memory Revolution Will Change Everything*. Penguin Books, 2009.
5. M. Dodge and R. Kitchin. "outlines of a world coming into existence": Pervasive computing and the ethics of forgetting. *Environment and Planning B*, 34(3):431–445, 2007.

6. A. R. Doherty, C. J. Moulin, and A. F. Smeaton. Automatically assisting human memory: A SenseCam browser. *Memory*, 7(19):785–795, 2011.
7. J. Gehrke, M. Hay, E. Lui, and R. Pass. Crowd-blending privacy. In *Advances in Cryptology - CRYPTO 2012 - 32nd Annual CryptologyConference*, volume 7417 of *Lecture Notes in Computer Science*, pages 479–496. Springer, 2012.
8. C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and Trends in Information Retrieval*, 8(1):1–125, 2014.
9. P. Halvorsen, S. Sægrov, A. Mortensen, D. K. Kristensen, A. Eichhorn, M. Stenhaug, S. Dahl, H. K. Stensland, V. R. Gaddam, C. Griwodz, and D. Johansen. Bagadus: An integrated system for arena sports analytics—a soccer case study. In *proc. of ACM MMSys*, pages 48–59, Mar. 2013.
10. S. Hodges, E. Berry, and K. Wood. SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory. *Memory*, 7(19):685–696, 2011.
11. B. K. Jacobsen, A. E. Eggen, E. B. Mathiesen, T. Wilsgaard, and I. Njølstad. Cohort profile: The Tromsø study. *International Journal of Epidemiology*, 2011.
12. D. Johansen, P. Halvorsen, H. Johansen, H. Riiser, C. Gurrin, B. Olstad, C. Griwodz, Å. Kvalnes, J. Hurley, and T. Kupka. Search-based composition, streaming and playback of video archive content. *Multimedia Tools and Applications*, 61(2):419–445, Nov. 2012.
13. D. Johansen, M. Stenhaug, R. B. A. Hansen, A. Christensen, and P.-M. Høgmo. Muithu: Smaller footprint, potentially larger imprint. In *proc. of the 7th IEEE International Conference on Digital Information Managemen*, pages 205–214, Aug. 2012.
14. H. D. Johansen, S. A. Pettersen, P. Halvorsen, and D. Johansen. Combining video and player telemetry for evidence-based decisions in soccer. In *proc. of the 1st International Congress on Sports Science Research and Technology Support*. INSTICC, Sept. 2013. Special Session on Performance Analysis in Soccer.
15. S. Kumpulainen, K. Järvelin, S. Serola, A. R. Doherty, A. F. Smeaton, D. Byrne, and G. J. F. Jones. Data collection methods for task-based information access in molecular medicine. In *proc. of the International Workshop on Mobilizing Health Information to Support Healthcare-related Knowledge Work*, pages 1–10, 2009.
16. M. L. Lee and A. K. Dey. Using lifelogging to support recollection for people with episodic memory impairment and their caregivers. In *proc. of the 2nd International Workshop on Systems and Networking Support for Health Care and Assisted Living Environments*, HealthNet '08, pages 14:1–14:3. ACM, 2008.
17. J. Machajdik, A. Hanbury, A. Garz, and R. Sablatnig. Affective computing for wearable diary and lifelogging systems: An overview. In *Machine Vision-Research for High Quality Processes and Products-35th Workshop of the Austrian Association for Pattern Recognition*. Austrian Computer Society, 2011.
18. J. Meyer, S. Simske, K. A. Siek, C. G. Gurrin, and H. Hermens. Beyond quantified self: Data for wellbeing. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 95–98, New York, NY, USA, 2014. ACM.
19. S. Spiekermann. The challenges of privacy by design. *Commun. ACM*, 55(7):38–40, July 2012.
20. I. Steadman. IBM's Watson is better at diagnosing cancer than human doctors. Technical report, Wired magasine, `http://www.wired.co.uk`, Feb. 2013.