AI & SOCIETY
https://doi.org/10.1007/s00146-024-01998-w

MAIN PAPER

Doubt or punish: on algorithmic pre-emption in acute psychiatry

Chiara Carboni[1] · Rik Wehrens[2] · Romke van der Veen[3] · Antoinette de Bont[4]

Received: 28 March 2024 / Accepted: 10 June 2024

Abstract
Machine learning algorithms have begun to enter clinical settings traditionally resistant to digitalisation, such as psychiatry. This raises questions around how algorithms will be incorporated in professionals' practices, and with what implications for care provision. This paper addresses such questions by examining the pilot of an algorithm for the prediction of inpatient violence in two acute psychiatric clinics in the Netherlands. Violence is a prominent risk in acute psychiatry, and professional sensemaking, corrective measures (such as patient isolation and sedation), and quantification instruments (such as the Brøset Violence Checklist, henceforth BVC) have previously been developed to deal with it. We juxtapose the different ways in which psychiatric nurses, the BVC, and algorithmic scores navigate assessments of the potential of future inpatient violence. We find that nurses approach violence assessment with an attitude of doubt and precaution: they aim to understand warning signs and probe alternative explanations to them, so as not to punish patients when not necessary. Being in charge of quantitative capture, they incorporate this attitude of doubt in the BVC scores. Conversely, the algorithmic risk scores import a logic of pre-emption into the clinic: they attempt to flag targets before warning signs manifests and are noticed by nurses. Pre-emption translates into punitive attitudes towards patients, to which nurses refuse to subscribe. During the pilots, nurses solely engage with algorithmic scores by attempting to reinstate doubt in them. We argue that pre-emption can hardly be incorporated into professional decision-making without importing punitive attitudes. As such, algorithmic outputs targeting ethically laden instances of decision-making are a cause for academic and political concern.

Keywords Machine learning · Acute psychiatry · Violence prediction · Nursing · Risk · Doubt

The relationship between (healthcare) professionals and emerging artificial intelligence (AI) tools has become central in contemporary academic and public debates. Questions concern the ways in which these technologies are likely to replace professionals (Chockley and Emanuel 2016), take over their tasks (Wong et al. 2019), alter the scope of professional decision-making and the nature of their work (Bullock 2019; Chan and Siegel 2019) or professionals' ability to domesticate these tools in practice (Avnoon and Oliver 2023; Topol Review 2019). These questions are usually discussed speculatively, and appear to be anchored in a future in which the use of novel technologies would have normalised. However, regardless of the slow pace of their implementation (Koutsouleris et al. 2022), AI tools are entering clinical practice along less formal routes, such as pilots or living labs (Archibald et al. 2021). This phenomenon also spans fields traditionally resistant to digitalisation, such as psychiatry (Bourla et al. 2018; May et al. 2001; Pickersgill 2018), where AI tools are being mobilised, for instance, to predict mood shifts in people with bipolar disorder (Semel 2021), classify, and predict behaviour in patients (Fernandes et al. 2017; Mulinari 2023), predict suicide attempts (D'Hotman and Loh 2020) or violent incidents (Borger et al. 2022).

If analyses of AI tools in practice are still few and far between (Jaton and Sormani 2023), empirical studies have begun to show the complexity inherent to processes of embedding AI systems in professional decision-making. Studies have shown, for instance, how AI can introduce new ambiguities in routine decision-making (Lebovitz 2019) and

✉ Chiara Carboni
chiara.carboni@tu-dresden.de

[1] Faculty of Linguistics, Literature and Cultural Studies, Technische Universität Dresden, Dresden, Germany

[2] Erasmus School of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] Erasmus School of Social and Behavioral Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands

[4] Tilburg School of Social and Behavioral Sciences, Tilburg University, Tilburg, The Netherlands

Published online: 27 June 2024

to their selective reliance on algorithmic outputs (Maiers 2017). This scholarship nuances our understanding of the ways in which algorithmic outputs are reshaping clinical practice in ways that exceed dominant narratives. Simultaneously, it alerts us to the potential marginalisation of other forms of knowing currently present in clinical practice, which, in times of crisis and workforce shortages, might end up sacrificed at the altar of algorithmically achieved efficiency (Henriksen and Bechmann 2020; Maiers 2017; Russell 2012; Schwennesen 2019).

Contributing to this emergent strand of scholarship, in this paper, we analyse the pilot of an algorithm for the prediction of inpatient violence[1] in two Dutch acute psychiatry clinics. By analysing nurses' reports in the electronic health record (EHR) and identifying 'predictive terms,' the algorithm was meant to provide a risk score for individual patients, thus flagging which patients warranted increased professional attention. In acute psychiatry, professionals' decisions are not only clinically relevant, but sanction coercive measures that can have direct consequences for patients' freedom. Zooming in on a pilot enables us to tease out the tensions between pre-existing practices around violence and emerging algorithmic logics. Attending to the different ways in which violence is constituted as an object for intervention, respectively, by the algorithmic output signal and by the local psychiatric nurses, this paper thus centres on the question of the implications of algorithmic outputs for professional decision-making. Specifically, we focus on how, and through which kinds of information, different agents (nurses, quantification instruments, and algorithmic risk scores) construct inpatient violence and on the implication of these different constructions in suggesting specific kinds of interventions.

In what follows, we first detail our theoretical footholds. To illuminate organisational and professional techniques for dealing with future risk, we combine work on future-oriented knowing and acting (Clarke 2016; Star 1991) with examinations of quantifications' and algorithms' ethics and performativity. By dwelling on reflections on the madness on decisions and its obfuscation in algorithmic logics (Amoore 2020; Derrida 2001), we come to characterize decision-making, particularly in acute clinical settings, as a moment of ethical deliberation, which algorithmic predictions attempt to obliterate. After laying out our methodology and describing our case more in depth, we turn to our empirical material, analysing, respectively, psychiatric nurses' practices around recognising, understanding, and dealing with inpatient violence, their use of quantification instruments, and the way the piloted algorithm attempts to intervene in nurses' practices.

1 Theoretical footholds

1.1 Future-orientedness

Future orientation constitutes a core aspect of many areas of social and organisational life (Flyverbom and Garsten 2021). In the context of computer-supported cooperative work, Adele Clarke (2016) has offered the concept of anticipation work, a type of invisible work engaging with the future as a space for (professional) action, and needed 'to optimize and live in preparation' (90; cf. also Star 1991; Strauss 1988). Clarke describes anticipation work as made up of three components: abduction, simplification, and hope. Abduction, a notion she borrows from Peirce, happens especially in conditions of 'genuine doubt or uncertainty or fear or great pressure to act' (quoted in Clarke 2016: 92; emphasis in original). It entails collecting empirical information and producing theories about it in an ongoing, yet tentative, way. Because of the inherent uncertainty of the conditions under which it is performed, all hypotheses generated abductively are adopted 'on probation' (91) and are thus always open for reconsideration. Simplification works by setting the boundaries and managing the complexity of the situation engaged. Finally, 'anticipation comes pre-wrapped in affect—hopefully inflected' (97)—that is, anticipation work is affectively laden and fuelled by a belief that the future might (be made to) be better than the present—at least for some. In our reading, anticipation work is also ethically laden, and ethics manifest in the probatory nature of hypotheses and theorisations anticipation work produces, which acknowledge doubts and the partiality of any knowledge.

If anticipation work helps us think through future-oriented epistemic practices, this type of abductive knowing concretises in action-oriented decisions. In this sense, Anderson (2010) proposes a typology of anticipatory action to flesh out the practices in the here and now 'paradoxically' justified by a (possible) future (778). Like anticipation work, anticipatory action emerges in conditions of uncertainty. Clarke's notion of hope becomes here a matter of staving off undesired futures, making sure 'that no bad surprises happen' (782). Anderson identifies three forms of anticipatory action: pre-emption (acting to neutralize threats that are yet to emerge), precaution (intervening before a perceived threat reaches a point of irreversibility), and preparedness (preparing for the aftermath of an event). In the context of acute psychiatry, precaution might mean increasing the sedation of a patient who has been raising their voice; pre-emptive

[1] In this paper, we use the term 'violence' to indicate a behaviour resulting in an incident (i.e., physical attacks to objects or people, or clear verbal attacks). With the term 'aggressivity,' we refer to behaviour that can be identified as warning signs for violence (e.g., restlessness, tone of voice, inappropriate language).

action would entail sedating a patient that an algorithm flagged as at risk of violence; preparedness entails making sure that sedatives are available on the ward.

Anticipation work is ongoing in acute psychiatric wards. Anticipation work, however, needs to consolidate into a decision—a moment that has been described as having a uniquely ethical character. Louise Amoore (2020) has recently revisited Derrida's (2001) reflection on the madness inherent to all forms of deciding in the wake of algorithmic decision-making. Since both the future and the implications of any decision are unknowable, Amoore argues that any form of anticipatory action is by nature a moment of ethical deliberation that exceeds purely epistemic consideration. As she states, 'to decide is to confront the impossibility of the resolution of difficulty; it is madness in the specific sense that it has no unified grounds' (Amoore 2020: 112). Though we return to Amoore's work later in this section, her discussion of the madness of decisions enables us to point out that anticipatory action entails ethical deliberation and requires ways to be response-able for the ramifications of one's decisions (Suchman 2023). Following Amoore, crucial aspects of such response-ability lie in foregrounding the doubt that spurs attempts to trace and open up alternative accounts, as well as and being cognizant of futures foreclosed by any human and algorithmic decisions.

In what follows, we work towards a framework aimed at embedding different risk prediction instruments currently used in psychiatric practice within a theory of anticipatory action. Guided by our empirical material, we focus mainly on two forms of risk prediction, namely, quantification-based risk prediction, and algorithmically enabled pre-emption.

1.2 Quantification and capture

Risk assessment instruments in psychiatry quantify aggressivity, providing nurses with checklists to score behavioural expressions presumed to be predictive of violence (see also Sect. 4.1). Quantification, that is, the 'production and communication of numbers' (Espeland and Stevens 2008: 401), has generated far-ranging sociological interest (Popp Berman and Hirschman 2018). Islam (2022) breaks down the phenomenon of quantification into processes of capture, specification, and appropriation. Processes of capture are central to how aggressivity is made into a graspable object in psychiatric clinics. Defined as the 'process of objectifying [a] social phenomenon so as to express it as a numerical quantity', capture often entails 'high levels of processing, manipulation, or abstraction' of aspects of social life (199). Moreover, quantitative capture decontextualizes one aspect of a lived experience in flux, and simplifies by decontextualising it. By attributing value to the numerical expression of knowledge and experience, capture thus risks dismissing more complex and relational accounts.

Forms of quantification can be seen as integral to the simplification component of anticipation work, in which it participates by performing the 'empirical,' by simplifying and producing what reality 'is.' Thus, albeit not future-oriented per se, quantification techniques can be integrated into forms of anticipation work. Because of this lack of inherent orientation to the future, however, we can speculate that quantification, in and of itself, does not suggest relating to the future through either precaution, pre-emption, or preparedness, but can be mobilised by actors as part of any type of anticipatory action. Moreover, the numbers that quantification instruments produce have, in themselves, no direct claim to the future. These numbers enact objects (e.g., aggressive patients) that are supposed to have existed in the past, and perhaps in the present. As we detail in the next section, this lack of claims to the future is a major way in which quantification instruments diverge from algorithms.

1.3 Algorithmic pre-emption

Unlike quantification instruments, machine learning algorithms can be seen as engaging in anticipation work. Analyses of computational technologies have emphasised the experimental way in which algorithms engage with reality. For Amoore (2020), the experimental nature of algorithms has to do with their ability to fine-tune and adjust parameters in an ongoing manner, thus subscribing to an eternally shifting version of the truth. Similarly, for Luciana Parisi (2019), in algorithmic engagements with the world, contingency and fallibility become productive forces in machine learning: algorithms learn through trial and error, generatively incorporating their own failures and thus engaging in an ongoing mode of optimisation that can be deferred ad infinituum (Halpern and Mitchell 2022).

Though both Amoore and Parisi chiefly discuss deep learning, the points they raise around the nature of learning apply to a broader class of machine learning algorithms. For our purposes, we can distil two ideas from these discussions. First, as Parisi argues, these algorithms perform their own particular abduction, adopting hypotheses on probation, and remaining open to learning from them when they prove wrong. Second, as Amoore points out, the truth against which these hypotheses are tested is removed from the lived world and anchored in a ground truth data set. This entails that algorithmic outputs cannot be judged as 'true' or 'false,' but must be seen as a (tuneable) function of their 'probabilistic proximity to, or distance from, a ground truth' (2020: 136).

If algorithmic abduction resembles other forms of anticipation work, algorithmic output represents a unique form of anticipatory action. As Amoore (2013) argues, algorithms' relating to the future, especially when mobilise to intervene on some form of risk, is fundamentally pre-emptive. That

is, by generating novel targets for action, algorithms tend to be geared towards creating the possibilities for action on yet unknown (or even still non-existing) threats. In Amoore's view, creating (and justifying) the possibility to act on a threat matters more, in an algorithmic logic, than the actual materialising of the threat itself. Rather than merely offering a representation of reality, thus, algorithms open up a space of action—and they do so through the generation of a univocal output signal. As Amoore puts it,

> among the most significant harms of contemporary decision-making algorithms is that they deny and disavow the madness that haunts all decisions. To be responsible, a decision must be made in recognition that its full effects and consequences cannot be known in advance. (2020: 120).

It is precisely in this effacing of the myriad of doubts and alternative explanations that they engaged in during processes of learning that algorithms fail us at the ethical level. Presenting a technically optimised output obfuscates the ethical nature of anticipatory action. Even more, algorithms propose pre-emption as the optimal future orientation, assuming the possibility of a future in which risk and ethical difficulties can be computationally. What is problematic here is not only the replacing of ethical doubts with the stochastic management of uncertainty, but also the lack of accountability for alternative futures that the univocality of algorithmic output (and of fantasies of optimisation) forecloses. The doubt that we teased out, above, as constitutive of the ethics of decision, is here effaced. In discussing possibilities for ethics in the wake of algorithmic decision making, Amoore (2020) suggests reinstating doubt at every point of machine learning's engagement with the world: from data, to ground truths, to data sets, to outputs. In our analysis, we follow this doubt as it differentially manifests, and sometimes is dealt with, in different practices around aggressivity and violence on acute psychiatric wards.

2 Settings and data

This paper builds on the ethnographic study of a 3-month long pilot that took place in two acute care clinics in a Dutch psychiatric hospital, to which we refer as clinic 1 and clinic 2. Clinic 1 is a 'high-intensity care' (HIC) clinic made up of 3 wards sharing 4 isolation cells. Each ward has between 8 and 10 rooms with en-suite bathrooms, as well as a communal living and dining room, and indoors smoking rooms. As a HIC clinic, clinic 1 receives all the involuntary admissions from the area—people brought in by ambulances or police after some public incident. More rarely, patients admit themselves. Although 'voluntaries' are, in principle, able to leave whenever they want, the clinic's personnel are liable for discharging them. This means that, when staff suspects patients could be at risk of suicide or violence, they can apply for a 'care authorisation' (zorgmachtiging), which revokes the patient's right to discharge themselves, forcing them to receive care. This is a long legal process in which lawyers and external psychiatrists are involved. In case of emergency situations (e.g., acute admissions), 'crisis measures' (crisismaatregelen) are usually granted, which, however, only last three working days (VWS n.d.). In clinic 2, the algorithm was piloted in a 'medium-to-high-intensity' ward, meant for patients who display less aggressivity or resistance to medications, but who are considered not stable enough to be discharged yet. This ward houses more patients, and has smaller rooms and shared bathrooms. Legal provisions and coercive measures in force in clinic 1 also apply to patients here.

Both clinics work with a system of 'freedoms:' as they stabilize, patients are allowed to do progressively more things: take walks on the hospital's terrain with a supervisor, go get their own groceries, go to the hairdresser, or go home for the weekend. Freedoms can always be revoked—for instance, when patients come back from a weekend away with a positive drug test. Because of the constant possibility of new admissions, particularly on weekends, when substance abuse is likelier, clinical staff is always under pressure to discharge patients, thus freeing beds. Clinics are usually staffed by one psychiatrist, one resident doctor, and several nurses (one to three per ward). Because of personnel shortages, flex workers are also called in every day. These people are often younger and have generic backgrounds as social workers, so they generally have little to no experience with this patient group and in dealing with aggressivity.

The pilot itself revolved around the introduction of an algorithm, developed within the same organisation, and aimed at predicting and pre-empting inpatient violence (see Sect. 4.2 for more details). Early intervention on patients' aggressivity emerged in Dutch policymaking in the early 2000s in the context of an international trend against coercive care in psychiatry. Whereas the general movement targeted coercion in general (i.e., involuntary medication, seclusion, and restraint, cf. Smith et al. 2022; WHO 2019), the Dutch programme, running from 2006 to 2012, aimed at phasing out the use of isolation cells specifically (Steinert et al. 2014). This resulted in the establishment of HIC clinics like clinic 1. However, programme fell short of its targets, and rates of isolation cell use raised again after its end. Even more significantly for our study, the reduction in rates of seclusion during the programme appears to have been reached through an increase in rates of involuntary medication (Vruwink et al. 2012). In other words, patients were likely being isolated less, but sedated more. Moreover, rates of sedation appear to have continued increasing in recent years (Van Melle et al. 2020). The piloted algorithm thus

appears to align with previous attempts at intervening on aggressivity before it escalates into violence—that is, before patients are considered in need of isolation.

Carboni had the opportunity to follow the pilot since its inception. She conducted 50 h of observations in both clinics, with the aim of observing how the risk scores produced by the algorithm would be discussed and used in clinical practice. She attended daily nurse–doctor handovers, as well as meetings in which wards' capacity and logistics were discussed. Nurses and psychiatrists also invited her to join their daily morning rounds to the isolation cells and, in the first days of fieldwork, some patient consultations. However, she stopped attending the latter after realising that, although medical students would also be in the room, her presence as a silent observer tended to distract patients, thus potentially impacting the consultation.

Instead, she started spending the shift in the nurses' station. Unlike psychiatrists' offices, which are removed from the ward, nurses' stations are either next to or inside the wards. They also have monitors showing isolated patients in real time. Conducting observations in these settings allowed her to be static in clinics in which most doors need to be locked, thus disrupting care practices as little as possible. Nurses' stations are also the place where the most nurse–patient interactions take place, as well as data registration. Being situated there, she was able to observe and discuss nurses' care and data practices, as well as their thoughts about difficult patients, organisational dynamics, aggressivity, and algorithms.

Carboni also kept in close contact with the data scientist leading the pilot. She had regular conversations with her about her ethnographic observations, and about the rationale behind various design and implementation choices. Carboni also conducted one 2-h interview with her, discussing more in depth the development and architecture of the model, as well as the lessons she had learned from the pilot. The data scientist organised two mid-term evaluations (one for each clinic) and three final evaluations of the pilot (one for each clinic, plus one with the patient council of the conglomerate the hospital belongs to). During these evaluations, professionals discussed their thoughts about algorithmic prediction and inpatient violence. Carboni attended all evaluations, and presented some insights from her fieldwork during the latter.

Bringing up issues of non-use during these meetings—as she had previously done in informal conversations—was met with somewhat defensive attitudes, particularly by one senior psychiatrist. Nurses did not object to her observations. After the end of the pilot, Wehrens, together with the data scientist and another doctor involved in the algorithm's development, organised three 1-h focus groups (with professionals, managers, and patients, respectively) to share further results from Carboni's analysis and discuss their views around the Brøset violence checklist (BVC, see Sect. 4) and its automation more in depth. Although they do not constitute the core of this analysis, the results of these focus groups have informed our thinking in developing this paper. The rest of the material (fieldnotes, observations from the evaluation meetings, and interview transcript) were coded abductively with the software Atlas.ti. Emerging themes, such as nurses' affective and embodied practices around aggressivity, dislike of the BVC, and playful attitudes towards algorithmic outputs are at the core of our analysis.

3 Violence: understanding and doubting

Psychiatric nurses on acute wards find themselves having to anticipate or deal with aggressivity many times a day. Since their stations are located in the wards, nurses are the professionals who interact most extensively with the clinics' patients. They are responsible for administering medications and executing routine drug tests; they also make sure all patients get enough food, organize discharges, and respond to the many requests coming from patients. Although only psychiatrists have formal consultations with patients, nurses informally talk to them throughout the shift, and have a '5-min contact' round every morning, to get a sense of patients' conditions.

Partially due to the topic of the pilot, aggressivity and violence on the ward were a common theme in our informal conversations with nurses. Recognising and dealing with signs of impending violence constitute a central component of nurses' work. Especially, in a clinical context in which many patients are psychotic, sings of aggressivity cannot be assessed in universal or absolute terms. Instead, assessing aggressivity entails picking up on subtle embodied signs that signal deviations from an individually defined 'normal:' 'Does the patient breathe more quickly, or more deeply than usual? Do they look at you askance? Do their muscles appear tense?' As nurses often emphasised in our conversations, 'what isn't threatening in one patient could very well be a threat in another one.' For instance, raising one's voice could qualify as aggressive behaviour in one patient, but only indicate hypoglycaemia in a diabetic one, and thus amount to an isolated incident that does not warrant disciplinary intervention—perhaps only a snack.

The example of the diabetic patient highlights how nurses' approach to the risk of violence entails attempting to understand the underlying causes of behavioural expressions that might be perceived as aggressive. For instance, nurses recount how there might be deeper emotional causes (e.g., being scared or sad) behind aggressive behaviour: 'if you address that, also the aggressivity goes away.' Taking signs of aggressivity at face value, rather than trying to understand their cause, thus emerges, first, as epistemically inadequate, because it effaces a multitude of alternative explanations by

assuming a univocal link between an internal (emotional and physiological) state and its external behavioural manifestation.[2] Moreover, and crucially, it is also ethically problematic, in that it would suggest violent interventions (sedation, isolation) that would be likely to escalate the situation.

The doubt mushrooming across nurses' accounts of how they assess the link between observed aggressivity and its cause is ethically charged and rooted in the experienced fallibility of attempts to understanding and relating to an object volatile yet threatening, such as inpatient violence. In their disciplinary power, nurses have to assess under which circumstances aggressive behaviour might be justified, and when it should be reined in. Distinguishing aggressivity as a risk factor for inpatient violence from aggressivity stemming from emotions that patients have a right to experience is as a tricky tightrope professionals have to walk in their decision-making. Professionals usually ignore the aggressivity displayed by patients who appeared 'quick to anger' as long as they are able to quickly apologise for their behaviour. Even when expressed in confrontational ways, anger does not necessarily warrant escalation on the part of professionals:

> In the second isolation cell there's a woman who has been moved here because of some violent behaviour. … [when the nurse opens the door], the patient appears pretty calm. … At the end of the conversation, when the psychiatrist offers her medications, the woman gets upset. She starts crying and tells them that she doesn't want them and doesn't need them. Nonetheless, she opens her hand and receives the two white pills a nurse hands her. Kneeling down, the nurse also passes her a paper cup with some water. The woman takes a small sip and then flings the cup to the ground floor, spilling all the water on the nurse's foot. This brings the conversation to a halt, and everyone leaves the cell, locking the woman inside. But even after this incident …, the psychiatrist states that the patient appeared much calmer. The staff agrees that they are going to see how things evolve in the evening, and possibly move her back to the ward the following day.

Although throwing a cup is punished by interrupting the conversation and leaving the patient in the cell, the small incident does not affect the psychiatrist's overall positive assessment—to the extent that she suggests trying to move the patient back to the ward. A rebellion does not qualify as aggressivity worth intervening on. In a similar way, signs of aggressive behaviour are also considered not worthy of intervention if another relational justification for them could be identified: sometimes, patients are upset by relatives' visits, or by tensions within the patient group. Appraising aggressivity in its relational dimension thus enables nurses to understand its causes through contextualisation efforts.

At the start of the pilot, nurses had agreed to consider (and thus start reporting) as violent episodes verbal threats and physical attacks to people and objects. However, it was not uncommon for nurses to ignore acts that might reasonably fall into one of these categories. For instance, one day Ralph, one of the nurses, came back to the nurses' station saying that a newly admitted patient had just spat in his face. 'So gross!' he commented, after sitting down at his desk and continuing his administrative work. The point here is not just that nurses working in acute settings might become desensitised to instances of violence, particularly when they do not put them at risk of physical harm. Our point is also not simply that they have an interest to work with as narrow a definition of violent incident as possible to save themselves some administrative burden. Rather, we argue that working with such a narrow definition, and thus de facto ignoring episodes that could be read as violent incidents, has to do with how ethically charged professionals perceived their violence-related interventions to be. This emerged during one of the first handovers we observed:

> When, at the end of the handover, I ask the nurse and the doctor about how they deal with aggression, they explain that what counts as a 'significant episode' is very complicated: 'If someone hits the window, like today, is it aggression to objects, or is it just that they got a bit angry?' Josh says. Karin, the doctor, agrees: 'Do you come in and pump them full of lorazepam?' 'Exactly, or put them in a straitjacket?'

The complexity of the negotiations nurses and physicians engage in when dealing with aggressivity is striking. These negotiations are made complex by, on one hand, the epistemic doubt surrounding attempts at classifying what they are seeing and, on the other, the ethical doubt that is about deciding whether interventions that are in themselves violent in their infringing on patients' autonomy, either at the physical or the neurochemical level, are warranted. Indeed, when nurses flag a patient as at risk of violent behaviour to a psychiatrist, this can set off a chain of measures that can be considered as violent in themselves: for instance, injecting medications in the bodies of patients who refuse to take them orally (an operation that requires several nurses to immobilize the patient in question), or dissolving them in their food; increasing their sedation levels; or, in the more extreme cases, locking them in an isolation cell.

Professionals' discussions of violence and aggressivity are guided by an awareness of their own fallibility, by the doubt inherent to any attempt to understand a complex reality—doubt stemming from the potential multiplicity

[2] Elsewhere, we dwell on the affective labour underpinning the knowledge- and decision-making of psychiatric nurses (cf. Egher, Carboni and Wehrens forthcoming 2024).

of explanations underpinning any behavioural expression they observe. This doubt materializes in the rejection of universal warning signs (replaced by knowledge of individuals' personal and clinical histories); in their probing of counterfactual explanations for aggressive behaviour (e.g., other feelings); and in their attempts at finding contextualising explanations beyond the individual that might justify aggressivity. We suggest that this is warranted, because these decisions are perceived as an ethical moment by nurses themselves.

4 Quantifying aggressivity, pre-empting violence

In most clinical settings, nurses are tasked with registering the bulk of clinical data (e.g., Islind et al. 2019). As explained above, in the clinics, we observed, nurses spend their shifts on the ward, in close interaction with patients. They thus collect much more varied and extensive observational data than psychiatrists, who only interact with patients during consultations. Throughout the shift, one of the ward's nurses writes up their observations for each patient in a Word document, which, before the end of the shift, is both copy-pasted in the EHR and printed out to be brought to the handover. These reports are relatively standardised. First, they describe the 5-min contact round in reasonably standardised language (e.g., 'confused,' 'friendly,' 'psychotic,' and 'verbally hyper but can be reasonably corrected'). Following this, they succinctly write up any episode worth of attention (bijzonderheden), which is usually articulated more in depth orally during handovers. This more narrative part of the report is followed by two highly structured sections: 'risks' (gevaar), detailing the reason for admission or other points for attention (e.g., psychosis or suicidality), and, finally, the result of the currently deployed violence risk assessment instrument: the Brøset violence checklist (BVC) score.

In this section, we briefly describe the functioning of the BVC, its enactment of aggression as a quantifiable risk object, and the hesitations the use of this instrument brings up in nurses. We then move to the algorithm at the centre of the pilot we followed, and try to understand in what ways the risk object it creates, and the use scripted in its organisational embedding, might deviate from previous instruments of quantification.

4.1 The BVC: quantification and its discontents

Developed in the late 1990s, the BVC is a violence risk assessment instrument aiming to offer a rough estimate of the upcoming 12–24 h. It requires ward nurses to assess patients based on the presence or absence of six items: confusion, irritability, boisterousness, physical threats, verbal threats, and attacks to objects. Items are to be factored in only if they represent a deviation from what is to be considered normal for a patient, e.g., boisterousness for a patient known to be quiet. Each observed item adds one point to the final score. The final score (ranging from 0 to 6) should assist further decision-making. Almvik et al. (2000) suggest taking 'preventive measures' for scores 1–2, and activating 'plans for handling an attack' for scores 3 and higher, for instance, increasing sedation or isolating patients.

Similar to nurses' practices, the BVC turns violence into a risk object that can be intervened upon before its escalation through appropriate anticipatory action (Anderson 2010). Unlike nurses, however, it does so by establishing a linear relation between factors (behavioural expressions) that it identifies as predictive, and episodes of violence themselves. Observing one or more of these expressions directly translates into an increased risk. Arguably, the direct connection the BVC operates between these behaviours and violence bears ontological implications. First, identifying individual behaviour as a predictor of violence, the BVC roots its causes firmly in individuals. Second, operating within a precautionary logic and overtly disregarding 'the motivations behind violent incidents' (Linaker and Busch-Iversen 1995), the BVC casts observation of individually defined risk factors as a sufficient condition for anticipatory action (De La Fabián et al. 2023). Finally, since risk is taken to reside in individuals, the possible interventions that scores 3 and higher call for will also target the individual patient who expressed the problematic behaviour. On a more general level, moreover, the BVC also assumes both the risk factors it identifies and violence itself to be clear-cut, commonsensical categories that can be uncontroversially applied to categorise patient behaviour.

Perhaps unsurprisingly, given the clash between how nurses think about violence and how the BVC operates, nurses had a complicated relationship with the instruments. Some nurses refused to fill it in, while others described it in informal conversations as 'a box-ticking exercise … with a lot of copy-paste going on, especially when no significant episodes happen during a shift.' Both during informal conversations and more formal evaluations of the pilot, nurses objected to the awkward temporality of the BVC, which in their view has no predictive value:

> Lola claims that the BVC doesn't help nurses, because it always refers to a situation they have already dealt with. Once we sit down for lunch, Ralph also explains that nurses score it at the end of their shift. However, if a patient shows any behaviour that could possibly escalate, they act on it immediately. If a patient knocks some objects off a table, or threatens other people, nurses will probably give them some time out in their

room, or in some cases even in the isolation cell. By the time they score the BVC, the situation has often already resolved.

As we learn from studies of quantification (Islam 2022), the BVC simplifies and de-contextualises the complex phenomenon of patient behaviour. It requires nurses to score aggressivity as if it was an external entity that they themselves were not experiencing and interacting with. As we have learned, on the ward, observations of aggressivity are rarely divorced by interventions aimed at countering or deescalating it. As a result, a high BVC ends up describing, ex-post, episodes that have already triggered cautionary measures. As emerged during handovers, patients observed to be too restless are immediately given timeouts, and isolation cells are prepared as soon as patients started 'thinking and acting chaotically' and becoming unintelligible for nurses.

The somewhat artificial separation between life on the ward and its punctual, quantified representation in the BVC is, to a certain extent, bridged by nurses. Crucially, nurses are both involved in the embodied, affective experience of aggressivity and violence, and are supposed to be the motor of its datafication. If quantifying something as complex and intangible as aggressivity often proves challenging, it gives them nonetheless an opportunity to instil the doubt that characterizes their approach to aggressivity in the BVC data they produce:

> [During the mid-term evaluation] Tim, a nurse, recounts how they had a patient who was proving really difficult to handle; it was hard for the nurses to pin down exactly why—and the patient would score 0 every day, though he gave them a lot to do. He goes on to talk about another patient, who sometimes would kick a door, or throw an ashtray on the floor. 'That was more of a rebellion than aggressivity,' according to Tim. Regardless, this other patient would score a 6—a score that, according to protocol, mandates isolating patients. 'In this case we did give him medications, but we decided to go back and change the score. They were all separate incidents, after all.'

If Tim points out how separate acts of rebellion over the course of a shift do not necessarily add up to an increased risk of violence, something more radical emerges from the fieldnotes above about how nurses produce and adjust the numbers in the BVC. Tweaking scores appears justified in cases when observed aggressivity is either hard to articulate through pre-established categories, or when it is doubted as a predictor of violence. These are moments, as we saw above, in which epistemic and ethical doubt, respectively, guide nurses' anticipation work. We want to suggest here that these moments of tweaking amount to attempts to make BVC scores not so much a more accurate representation of reality, but a more accurate representation of nurses' doubt. By instilling doubt in the numbers they produce, nurses try to make BVC scores more internal to the life on the ward and to aggression as a relational, experienced phenomenon.

This matters, because, although barely brought up in handovers, nurses do report using BVC when 'it's busy and [they] have so little time and personnel,' and thus 'need to understand quickly what's going on.' Due to personnel shortages and the very nature of work on acute psychiatric wards, nurses sometimes benefit from oversimplified but quick ways to assess the situation. In what follows, we analyse the algorithm that was aimed at automating the BVC scoring, and trace how its differential relation to doubt and externality influenced its reception by nurses.

4.2 'Just a stupid outcome indicator:' training a violence prediction algorithm

Inpatient violence emerged as a target for the algorithm, partially because the organisation's data science department had access to a similar model.[3] The original model was intended to predict violence episodes in patients over a period of 2 weeks, and had previously been tested on data from the psychiatric hospital in which the pilot was run. The organisation's data science department thus decided to retrain it to predict the following 24 h—which was considered a more actionable timeframe. The algorithm was thus developed with the aim of replacing the BVC with an automated risk assessment, and of assisting professionals in their decision-making during handovers.

Emilia, the data scientist behind the repurposed algorithm, explained to us that the topic of violence had been selected by the organisation's management mostly because of the model's availability, and secondarily due to the clinical prevalence of violent episodes. From her perspective, however, violence was not an experienced threat, but rather an abstract entity—as she described, 'a bad outcome indicator.' To train the algorithm, Emilia used supervised machine learning techniques, such as random forest and bag of words, which teach the algorithm predictive terms based on their correlation with an indicator (i.e., a datafied proxy for violent episodes). In the clinics we studied, nurses were responsible for reporting violent incidents through an online portal. However, as mentioned, they rarely did. This 'bad outcome indicator' (i.e., a painfully incomplete data set) caused issues in the training process and, down the line, in the performance of the algorithm:

[3] Although research on this model has been published, we cannot reference it here for purposes of anonymity.

> [The algorithm] is not doing what it should do [because] we don't have the incidents and the almost-incidents. [So] the model says 'okay, I see words like 'aggressive' and … 'throwing' or 'hit'—and there is no incident. Huh? What do I do? It's weird.' And it *is* weird, because there was an incident, but it was not reported. (Emilia).

Concretely, the algorithm was trained on a data set correlating nurses' notes for each patient with presence or absence of a VIM report for that patient during that shift. Based on this, it identified a series of words that, supposedly, either increase or decrease the likelihood that that patient would initiate an incident. Predictive terms included terms such as 'aggressive,' 'reacts,' 'offered' (likely referring to medications), 'angry,' 'emergency medications,' and 'colleague' (likely referring to the fact that nurses always approach aggressive patients accompanied by at least a colleague). Being linked to behaviour and interactions, these terms are supposed to capture aspects of aggressivity that would be overlooked in traditional quantified risk assessments.

Here, we encounter a first aspect in which the algorithm diverges from the BVC. The focus on words used in nurses' reports seems to attempt to push the boundaries of quantification by spanning more subtle signs of aggressivity and relational causes of violence. However, the way it appraises nurses' reports is crucial. First, the algorithm assumes a linear relationship between the presence of predictive terms (and the behaviour and interactions they refer to) and the increase of violence risk. In other words, whenever the word 'colleague' is mentioned in a report, the risk score for that patient automatically goes up. This predictive logic, which it shares with the BVC, translates into performance issues. In Emilia's view, based on her perusal of the data set, 'there are people that … quite frankly just … have some kind of one-off explosion of violence.' Indeed, this resonates with nurses' experiences of some situations of tension resolving themselves and others being impossible to predict. Interestingly, in Emilia's view, only the latter ('false negatives') are an issue: what worries her is not being able to identify patients that do become violent, rather than the chance of flagging, and potentially intervening upon, patients that would not have become violent after all.

A second way in which the algorithm's appraisal of risk differs from practices on the ward concerns its identification of predictive terms. These terms emerged as 'predictive' based on techniques of data set creation, supervised machine learning and model tuning. If nurses can tweak the BVC to provide a better representation of their doubts, tweaking, (or tuning) happens here in the relation between data scientists and algorithm. Tuning is chiefly a matter of statistical accuracy, referred to the (shaky) ground truth of violence reports, rather than to the situated and embodied relations nurses have with a specific patient. This, of course, matters, in its externality to life on the ward, in that it ends up constituting what counts as risk in a way that is separated from the lived experience of that risk.

It is, of course, impossible to detail what tuning actually does when working with algorithms. What we can, however, look at, are the risk scores the algorithm generated throughout the pilot. As Emilia disclosed during the final evaluation, during those 3 months, the algorithm had flagged as at risk of violence (≥ 0.5) more than 500 cases (i.e., patients/day) that had been given a BVC of less than 3. In other words, the algorithm had suggested adopting violence-pre-empting measures on 500 occasions in which nurses would have not initiated any intervention. Risk scores are thus a major aspect in which the BVC diverges from the algorithm. Clearly, this has to do with the way the algorithm is tuned: we have seen how Emilia is worried about false negatives, about the algorithm possibly not picking up on all risk behaviour. The risk of false negatives, thus, drastically outweighs the reality of false positives.

4.3 Reinstating doubt in algorithmic predictions

Risk scores were shared with each clinic every morning, before the general handover, in the form of an Excel sheet listing all the clinic's patients' names and their risk scores (a number from 0 to 1), ranked from highest to lowest. In Emilia's and the managers' plans, professionals in the two clinics would discuss the aggression scores during the morning handovers, during which nurses share their reports with psychiatrists, and discuss which patients need a psychiatric consult or a new treatment plan.

If patients' aggressivity is a major topic of discussion during handovers, it is never a topic tackled by referring to BVC scores. Perhaps, unsurprisingly, thus, the algorithmically produced risks scores were not part of this decision-making throughout the whole pilot. Even in the first days, when the algorithm was still new and exciting for professionals, they never quite took seriously the knowledge it produced in their deliberations:

> The algorithm only came up towards the end of the meeting. Josh, the nurse, commented that it was weird that the patient with the 2nd and 4th highest scores were scoring so high. Especially for the 4th one, he commented that 'it's just a matter of personality, he spends a lot of time in his room.'

Although the algorithm was intended to support decision-making, it is interesting to notice how scores actually generate decisions by coming up with targets for intervention (cf. Amoore 2013; Ratner and Elmholdt 2023). If professionals are the ones to finally implement any decision, it is also true that a major way in which algorithmic scores are supposed

to contribute to decision-making is by deviating from professional judgements—that is, by flagging target that are not on their radar. However, as emerges from the quote above, whenever the algorithm diverged from their assessment of a patient's situation, professionals simply assumed it was wrong. Some of them even attempted to theorize why the algorithm would assess some patients wrongly:

> Josh explained how nurses are often surprised by some of the patients who had been scored as high-risk. He shared his theory about why: 'Sometimes [the algorithm] will score as high-risk someone who really isn't, because it picks up on the 'aggressive' in a sentence that actually says 'not aggressive.'' He showed me the file with the risk scores for that day, comparing it with the printout from the handover. He saw the name of the patient scored as the highest risk: 'This one, for example, doesn't make any sense. This is a patient who is asking to be isolated from the group. That has nothing to do with violence!'

Investigating whether the algorithm was getting things wrong is beyond our analytical scope. These attempts at reverse-engineering algorithmic reasoning show that professionals consider risk scores on a lower epistemic level than their assessment. Such attitude, which was the only way nurses and psychiatrists engaged with risk scores, starkly rejects algorithmic pre-emption and the assumption that machine learning might pick up on aggressivity before nurses.

During the pilot's final evaluation, Carl, the head of one of the clinics reflected that this mode of engagement likely stemmed from the gap between algorithmic and professional ways of assessing aggressivity: '[the algorithm] sees something in the [EHR] notes and [the score] goes up. And when we see it, we think—yeah, but of course, that's not aggressivity.' Himself a nurse, Carl questions here the linear connection between words used to describe patients' behaviour and violence risk. First, the words the algorithm considers 'predictive' are often generic (e.g., 'colleague,' or 'offered'). Second, what is registered in the EHR is a simplified, decontextualised version of life on the wards: nurses routinely add a wealth of contextualising details to their account orally during handovers. Since understanding the possible motivations behind a patient's behaviour is central to nurses' techniques around aggressivity, they consider the algorithm's sole reliance on EHR data insufficient to assess aggressivity.

Inspired by Amoore's (2020) articulation of ethics in the wake of machine learning as founded on a reinstating of doubts and alternatives against the certainty of output signals, we read nurses' rejection of the linear logics of prediction and insistence on contextualisation as an ethically laden attempt at opening up the multiplicities that algorithmic scores reduce to one. The only meaningful way to engage with a score's reductionism is then, as we saw Josh do above, trying to understand which relational dynamics the algorithm might be objectivising, which stories it might be ignoring, which alternative futures it might be pre-empting. Nurses are, as it were, reinstating doubt in algorithmic outputs. We elaborate on the theoretical and practical implications of this in our discussion.

5 Discussion: on doubt, futures, and data

In this paper, we have examined three different ways of relating to and dealing with aggressivity and violence risk: nurses care practices, the BVC (and nurses' interference with it), and the algorithmic risk scores. We have argued that nurses mobilize a kind of doubt that is simultaneously epistemic and ethical. Doubt troubles both the linearity between expressed sign of aggressivity and underlying states experienced by patients, and the opportunity of intervening on these signs (thus potentially escalating situations that might have resolved themselves). Nurses' practices illuminate how deciding entails dealing with a multiplicity of potential alternative accounts, and is, as such, an ethical moment. The probing of alternative explanations is central to professionals', and specifically nurses', practices around aggressivity and violence on acute psychiatric wards. In their interactions with patients, in their use of quantification instruments, and even in their relating to algorithmic risk scores, nurses embody an attitude of opening up possibilities, resisting to linear logics and insisting on a proliferation of possible alternative accounts. Indeed, it is this very opening up of alternatives that makes it possible for them to morally navigate heavily ethically charged decision-making moments. Because of this doubtful attitude, nurses' anticipatory action tends to be informed by precaution: they monitor the situation by attempting to notice warning signs, yet regard intervention as warranted only when it might prevent a perceived threat from escalating.

Due to their constant presence on the ward, nurses understand themselves as internal to the world they are appraising and attempting to intervene on. As such, they are aware of their need to account for the ramifications of their actions—or lack thereof. This internality and accountability are progressively displaced with instruments for quantification and risk pre-emption. The logic of precaution also applies to the BVC: although it crystallises life on the ward in strange temporalities, doubt can be effectively incorporated in the final BVC score. Nurses thus meaningfully shape quantitatively mediated anticipatory action. Conversely, the algorithm imports a pre-emptive orientation that is unrecognisable to nurses. Pre-emption means that the space for doubt is limited to the training process, and contained in the complex algorithmic architecture and its finetuning. Its univocal

output, by generating decontextualised risk scores that aim to efface doubt and alternative accounts, intervenes at the level of anticipatory action itself. Although professionals try to reinstate doubt in the score, the univocality of the output leaves them with a dichotomous option to either subscribe to or altogether reject the score. Moreover, the algorithm's pre-emption attempts to identify patterns that might be beyond professionals' awareness, thus generating targets for intervention that are not necessarily known (or, as it were, not yet existing).

If other applications of machine learning in acute psychiatry are possible, it is worth here taking seriously the implications of introducing pre-emptive orientations in clinical settings, like this particular algorithm (and others, e.g., Borger et al. 2022) attempts to do. This appears particularly urgent in acute clinical settings in which shortages of trained personnel translate into emphasised need for securitisation. First, algorithms in psychiatry have to face the problem of the limits of calculability (Amoore 2020): modelling objects that, like violence, are known to be volatile and hard to pin down. Even with machine learning, these objects continue to elude predictions (the reader might remember Emilia's comment on some patients showing 'explosions' of violence). When the objects of algorithmic outputs are ethically laden, like in the case of violence, we should interrogate the drive towards effacing their inherent difficulty through algorithmisation. In the case examined in this paper, we might argue that the difficulties emerging from this algorithm are not so much technical ones. Even if the algorithm had access to more and more diverse datapoints (e.g., physiological data related to aggressivity), it is its very attempt at addressing, and somehow simplifying, such an ethically charged decision-making moment that deserves critical scrutiny. It is worth considering what a similar case might tell us about how the desire to embrace data and machine learning is currently reshaping cultures of medicine—even in a traditionally digitalisation-resistant field such as psychiatry (e.g., Pickersgill 2018; Stevens et al. 2020). The pilot we examine here was driven by a desire to have an—admittedly painfully simple—algorithm partially automate a process that is known to be epistemically and ethically extremely complex. Arguably, this raises fundamental questions around what aims and desires the discourse on contemporary techniques of data analysis is making thinkable and utterable in the medical field.

Furthermore, noticeably, the pre-emptive orientation introduced by the algorithm examined here reshapes violence as an object of intervention in ways that warrant reflection. On the one hand, the algorithm's pre-emptive orientation enables professionals to imagine a less violent future for acute psychiatry. Nurses and psychiatrists often shared their unease with practices of patient isolation, which they recognised as problematically violent, and which Dutch policymakers have, as discussed above, attempted to phase out in the past (Van Melle et al. 2020). In this sense, catching threats before they emerge would allow them to phase out this controversial practice. Although pre-emption would likely entail more sedation, we can see that the introduction of risk scores is, at least partially, justified by a will to move acute psychiatry towards more ethical ways of dealing with unstable patients.[4]

On the other hand, algorithms' pre-emptive future orientation is a cause for concern, especially when they are meant to support decision-making in acute psychiatric settings.[5] If the trade-off between sensitivity and specificity (i.e., between false positives and false negatives) is a well-known issue in statistics, here it emerges starkly in its ethical import. Algorithmic risk scores presuppose linear relationships between signs of aggressivity and futures of violence; they recommend action by foreclosing the possibility for alternative interpretations of patients' behaviour, and for differential, doubtful ways of weighing particular episodes. Untouched by nurses epistemic and ethical doubts, the algorithm ends up displaying a drastically more punitive logic than nurses subscribe to. We can speculate that this might have to do with the model being tuned without nurses' involvement, and that a differential tuning of such a simple algorithm might have produced less strikingly off outputs. However, the point stands that, albeit attempting to recast aggressivity as relational and subtle through a focus on words, the algorithm falls dramatically short of considering the maybes, the doubt, the efforts towards contextualising and the search for alternative explanations that are essential parts of nurses' approach to violence. Moreover, its tuning to yield as few false negatives as possible fits with the general idea of security the algorithms is mobilised to achieve. When applied to issues such as violence algorithms' pre-emptive orientations are likely to enhance trajectories of increased securitisation that appear disproportionately punitive towards patients.

[4] Whether sedation amounts to a less violent measure is a point worth discussing, but beyond the scope of this paper (cf. Dumit 2012).

[5] As one of our anonymous reviewers pointed out, this does not even begin to address the question of whether a similar algorithm could be approved for use in clinical practice. Indeed, shifts in the regulatory landscape set off by the General Data Protection Regulation, the European Health Data Space and the AI Act complicate the question of how to ethically, and indeed legally, access and process sensitive patient data. In the case of the algorithm examined in this paper, the necessity of a legal assessment was dismissed because actors considered the pilot as falling within the scope of primary use of EHR data (i.e., clinical care). However, recent regulatory developments are likely to make thorough assessments a more pressing necessity in the future.

Indeed, the punitive logic that we have seen emerging in the final evaluation of the pilot, with risk scores' disproportionate number of 'false positives,' warrants reflection. As we have seen, integral to the algorithm is an assumption that it should identify risk nurses are not aware of, thus generating new targets for intervention. This is characteristic of pre-emptive approaches that, in the name of security, aim at obliterating threats before they emerge (Amoore 2013, 2020). Punishing (including sedating) patients earlier, and more, appears as the only way that an algorithm can live up to its promise of pre-emption in acute psychiatry. This, as we have shown, represents an attempted shift away from a field of clinical practice where decisions are acknowledged to be situated, tentative and ethically charged. Other machine-learning approaches to patient and personnel safety in acute psychiatry might be possible. However, as we suggest, they need to take into account, and indeed make central, the doubt at the heart of decision-making in these settings.

In closing, the informal ways through which algorithms with potentially serious implications are slipping from research settings, into pilots, into (potentially) clinical practice, albeit marginal to our analysis here, warrants at least mentioning. Like previous studies of machine learning in clinical practice (Lebovitz 2019; Maiers 2017), our engagement with this pilot has yielded a picture of selective reliance on algorithmic predictions. Nonetheless, it worth taking seriously the work that risk scores such as the ones analysed here might do in reshaping psychiatric practice in the future. If pre-emption is a feature of many applications of machine learning dealing with some sort of clinical risk, particularly in acute settings, machine learning in psychiatry necessarily impinges on behavioural and affective components that make its ethical stakes particularly pronounced. As we suggest, this dynamic, and the ease with which it seeps into practice, warrants careful observation, and perhaps precaution, within the academic community and beyond.

Acknowledgements We thank all the participants who shared their time and insight, contributing greatly to our research. We are also thankful to the TechSoc cluster at Copenhagen Business University, and in particular to Nanna Bonde Thylstrup, Mikkel Flyverbom and Frederik Schade, for providing generous and valuable comments that helped us improve our manuscript. This article, and the PhD project of which it is part, was made possible through the financial support of the Medical Delta program Journey from Prototype to Payment and of the Erasmus School of Health Policy and Management.

Data availability The data that support the findings of this study are stored according to the data management guidelines of Erasmus University Rotterdam. Due to the sensitive nature of some of these data, they are not publicly available. The data are, however, available from the authors upon reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Almvik R, Woods P, Rasmussen K (2000) The Brøset violence checklist: sensitivity, specificity, and interrater reliability. J Interpers Violence 15(12):1284–1296. https://doi.org/10.1177/088626000015012003

Amoore L (2013) The politics of possibility: risk and security beyond probability. Duke University Press, Durham and London

Amoore L (2020) Cloud ethics: algorithms and the attributes of ourselves and others. Duke University Press, Durham and London

Anderson B (2010) Preemption, precaution, preparedness: anticipatory action and future geographies. Prog Hum Geogr 34:777–798

Archibald M, Wiebe S, Rieger K, Linton J, Woodgate R (2021) Protocol for a systematic review of living labs in healthcare. BMJ Open 11(2):e039246

Avnoon N, Oliver AL (2023) Nothing new under the sun: Medical professional maintenance in the face of artificial intelligence's disruption. Big Data Soc. https://doi.org/10.1177/20539517231210269

Borger T, Mosteiro P, Kaya H, Rijcken E, Salah AA, Scheepers F, Spruit M (2022) Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting. Expert Syst Appl 199:116720

Bourla A, Ferreri F, Ogorzelec L, Peretti C, Guinchard C, Mouchabac S (2018) Psychiatrists' attitudes toward disruptive new technologies: mixed-methods study. JMIR Mental Health 5(4):e10240. https://doi.org/10.2196/10240

Bullock JB (2019) Artificial intelligence, discretion, and bureaucracy. Am Rev Public Adm 49(7):751–761. https://doi.org/10.1177/0275074019856123

Chan S, Siegel EL (2019) Will machine learning end the viability of radiology as a thriving medical specialty? Br J Radiol 91:20180416

Chockley K, Emanuel E (2016) The end of radiology? Three threats to the future practice of radiology. JACR 13(12PtA):1415–1420

Clarke A (2016) Anticipation work: abduction, simplification, hope. In: Bowker GC, Timmermans S, Clarke AE, Balka E (eds) Boundary objects and beyond: working with leigh star. MIT Press, Boston, pp 85–119

D'Hotman D, Loh E (2020) AI enabled suicide prediction tools: a qualitative narrative review. BMJ Health Care Inform. https://doi.org/10.1136/bmjhci-2020-100175

De La Fabián R, Jiménez-Molina Á, Pizarro Obaid F (2023) A critical analysis of digital phenotyping and the neuro-digital complex in psychiatry. Big Data Soc. https://doi.org/10.1177/20539517221149097

Derrida J (2001) Cogito and the history of madness. In: Bass A (ed) Writing and difference. Routledge, London, pp 31–63

Dumit J (2012) Drugs for life: how pharmaceutical companies define our health. Duke University Press, Durham and London

Espeland WN, Stevens ML (2008) A sociology of quantification. Eur J Sociol 49(3):401–436. https://doi.org/10.1017/S0003975609000150

Fernandes BS, Williams LM, Steiner J et al (2017) The new field of 'precision psychiatry.' BMC Med. https://doi.org/10.1186/s12916-017-0849-x

Flyverbom M, Garsten C (2021) Anticipation and organisation: seeing, knowing and governing futures. Organ Theory 2(3):26317877211020324

Halpern O, Mitchell R (2022) The smartness mandate. MIT Press, Cambridge

Islam G (2022) Business ethics and quantification: towards an ethics of numbers. J Bus Ethics 176(2):195–211

Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M (2022) From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health 4:e829–e840

Lebovitz S (2019) Diagnostic doubt and artificial intelligence: an inductive field study of radiology work. ICIS 2019 Proceedings 11

Linaker OM, Busch-Iversen H (1995) Predictors of imminent violence in psychiatric inpatients. Acta Psychiatr Scand 92(4):250–254

Maiers C (2017) Analytics in action: users and predictive data in the neonatal intensive care unit. Inf Commun Soc 20(6):915–929

May C, Gask L, Atkinson T, Ellis N, Mair F, Esmail A (2001) Resisting and promoting new technologies in clinical practice: the case of telepsychiatry. Soc Sci Med 52(12):1889–1901. https://doi.org/10.1016/S0277-9536(00)00305-1

Mulinari S (2023) Short-circuiting biology: digital phenotypes, digital biomarkers, and shifting gazes in psychiatry. Big Data Soc. https://doi.org/10.1177/20539517221145680

Parisi L (2019) Critical computation: digital automata and general artificial thinking. Theory Cult Soc 36(2):89–121

Pickersgill M (2018) Digitising psychiatry? Sociotechnical expectations, performative nominalism and biomedical virtue in (digital) psychiatric praxis. Sociol Health Illn 41:16–30. https://doi.org/10.1111/1467-9566.12811

Popp Berman E, Hirschman D (2018) The sociology of quantification: where are we now? Contemp Sociol 47(3):257–266

Ratner HF, Elmholdt K (2023) Algorithmic constructions of risk: anticipating uncertain futures in child protection services. Big Data Soc. https://doi.org/10.1177/20539517231186120

Semel BM (2021) Listening like a computer: attentional tensions and mechanised care in psychiatric digital phenotyping. Sci Technol Human Values. https://doi.org/10.1177/0162243921102637

Smith GM, Altenor A, Altenor RJ et al (2022) Effects of ending the use of seclusion and mechanical restraint in the Pennsylvania State Hospital System, 2011–2020. Psychiatr Serv 74(2):173–181. https://doi.org/10.1176/appi.ps.202200004

Star SL (1991) The sociology of the invisible: The primacy of work in the writings of Anselm Strauss. In: Maines DR (ed) Social organisation and social process: Essays in honor of Anslem Strauss. Aldine de Gruyter, Hawthorne, pp 265–283

Steinert T, Noorthoorn EO, Mulder CL (2014) The use of coercive interventions in mental health care in Germany and the Netherlands: a comparison of the developments in two neighboring countries. Front Public Health 2:141. https://doi.org/10.3389/fpubh.2014.00141

Stevens M, Wehrens R, de Bont A (2020) Epistemic virtues and data-driven dreams: on sameness and difference in the epistemic cultures of data science and psychiatry. Soc Sci Med 258:113116

Strauss AL (1988) The articulation of project work: an organisational process. Sociol Q 29:163–178

Suchman L (2023) The uncontroversial 'thingness' of AI. Big Data Soc. https://doi.org/10.1177/20539517231206794

Topol Review (2019) Preparing the healthcare workforce to deliver the digital future. Final Report February 2019–A call for evidence. Health Education England. https://topol.hee.nhs.uk. Accessed 11 Mar 2024

Van Melle AL, Noorthoorn EO, Widdershoven GAM et al (2020) Does high and intensive care reduce coercion? Association of HIC model fidelity to seclusion use in the Netherlands. BMC Psychiatry 20:469. https://doi.org/10.1186/s12888-020-02855-y

Vruwink FJ, Mulder CL, Noorthoorn EO et al (2012) The effects of a nationwide program to reduce seclusion in the Netherlands. BMC Psychiatry 12:231. https://doi.org/10.1186/1471-244X-12-231

Wong SH, Al-Hasani H, Alam Z et al (2019) Artificial intelligence in radiology: how will we be affected? Eur Radiol 29:141–143. https://doi.org/10.1007/s00330-018-5644-3

Egher C, Carboni C, Wehrens R (forthcoming) The role of affective labor in expertise: bringing emotions back into expert practices. Medicine Anthropology Theory.

Jaton F Sormani P (2023) Enabling 'AI'? The situated production of commensurabilities. Social Studies of Science 53(5): 625-634.

Henriksen A and Bechmann A (2020) Building truths in AI: Making predictive algorithms doable in healthcare. Information, Communication & Society 23(6): 802-816.

Russell B (2012) Professional call centres, professional workers and the paradox of the algorithm: The case of telenursing. Work, Employment and Society 26(2): 195–210.

Schwennesen N (2019) Algorithmic assemblages of care: Imaginaries, epistemologies and repair work. Sociology of Health & Illness 41(S1), pp. 176–192.

World Health Organization (WHO) (2019). Freedom from coercion, violence and abuse: WHO quality rights core training: Mental health and social services. Available at:https://iris.who.int/bitstream/handle/10665/329582/9789241516730-eng.pdf (last accessed: 17-06-2024).

Islind AS, Lindroth T, Lundin J and Steineck G (2019) Shift in translations: Data work with patient-generated health data in clinical practice. Health Informatics Journal 25(3): 577-586.