BMJ Open

# Study protocol for screening excessive gambling behaviors on French online gambling sites

| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2016-014600 |
| Article Type: | Protocol |
| Date Submitted by the Author: | 06-Oct-2016 |
| Complete List of Authors: | Perrot, Bastien; University of Nantes, EA 4275 SPHERE<br>Hardouin, Jean-Benoit; University of Nantes, EA 4275 SPHERE<br>Costes, Jean-Michel; Observatoire Des Jeux<br>Caillon, Julie; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders"<br>Grall-Bronnec, Marie; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders"<br>Challet-Bouju, Gaëlle; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders" |
| <b>Primary Subject Heading</b>: | Addiction |
| Secondary Subject Heading: | Public health |
| Keywords: | Gambling, Prevention, Latent class model, PUBLIC HEALTH, Online gambling |
| | |

# Study protocol for screening excessive gambling behaviors on French online gambling sites

Bastien PERROT[1*], Jean-Benoit HARDOUIN[1], Jean-Michel COSTES[2], Julie CAILLON[3,1], Marie GRALL-BRONNEC[3,1], Gaëlle CHALLET-BOUJU[3,1]

[*] Correspondence:
Bastien Perrot
bastien.perrot@univ-nantes.fr
EA 4275 - IRS2, 22 boulevard Bénoni Goullin – 44200 Nantes, France
Tel. +33 253009129

[1] EA 4275 SPHERE “methodS in Patient-centered outcomes and HEalth ResEarch”, Faculties of Medicine and Pharmaceutical Sciences, University of Nantes, France

[2] Observatoire des jeux, Ministère des Finances, Paris, France

[3] Clinical Investigation Unit BALANCED “BehaviorAL AddictioNs and ComplEx mood Disorders”, Department of Addictology and Psychiatry, University Hospital of Nantes, France

Word count: 3884

## ABSTRACT

### Introduction

Since the legalization of online gambling in France in 2010, gambling operators must implement responsible gambling measures to prevent excessive gambling practices. However, actually there is no screening procedure for identifying problematic gamblers. Moreover, although several studies have already been performed using several datasets from online gambling operators, the authors deplored several methodological and clinical limits that prevent scientifically validating the existence of problematic gambling behavior. The aim of this study is to develop a model for screening excessive gambling practices based on the gambling behaviors observed on French gambling websites, coupled with a clinical validation.

### Methods and analysis

The research is divided into three successive stages. All analyses will be performed for each major type of authorized online gambling in France (horse race betting, sports betting, poker and authorized lotteries). The first stage aims at defining a typology of French online gamblers based on their gambling behavior. This analysis will be based on the databases of the Authority for Regulating Online Gambling (ARJEL) and the Française Des Jeux (FDJ). For the second stage that will aim at determining a score to predict whether a gambler is problematic or not, we will cross answers from the Canadian Problem Gambling Index with real gambling data. The objective of the third stage is to validate clinically the score previously developed. Results from the screening model will be compared to the diagnosis obtained with a telephone clinical interview including diagnostic criteria for gambling addiction.

### Ethics and dissemination

This study was approved by the local Research Ethics Committee (GNEDS) on March 25, 2015. Results will be presented in national and international conferences, submitted to peer-reviewed journals and will be part of a PhD thesis. A final report with the study results will be presented to the ARJEL, especially the final screening model.

## Strengths and limitations of this study

- This study will contribute to setting up an innovative prevention measure to inform and protect gamblers as early as possible.
- Data will be representative of the French online gamblers' population; all types of online gambling will be included.
- Clinical validation of the final screening model.
- Use of advanced statistical methods.
- Potential selection bias at stages 2 and 3.
- Difficulty to include time-based gambling indicators.

## Trial registration

NCT02415296

## Keywords

Online gambling, problem gambling, prevention, classification, latent class model, predictive model

## INTRODUCTION

The majority of gamblers have a controlled and recreational gambling practice, but some of them lose control of it. In the Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders, (DSM-5), gambling disorder is defined as a "persistent and recurrent problematic gambling behavior leading to clinically significant impairment or distress" [1]. A French national survey conducted in 2015 indicated that 3.9% of past-year gamblers had gambling problems, including 0.9% of excessive gamblers [2]. With regard to online gambling only, the prevalence of gambling problems rises to 17%, including 6.6% of excessive gamblers [3]. Several factors can explain the fact that online gambling is more addictive than offline gambling; especially greater accessibility, increased disinhibition and higher event frequency [4].

The opening to competition and regulation of the French online gambling sector was provided by the law No 2010-476 of 12 May 2010. This law included the creation of the Regulatory Authority for Online Gambling (ARJEL in French). The ARJEL can issue licenses to online operators for only three types of games: horse race betting, sports betting and poker. Lotteries and scratch games provided by the Française des Jeux (FDJ in French) are also authorized online because of a particular waiver. The ARJEL compiles data from all accounts created on authorized online gambling sites in France, namely 1,352,000 active accounts in the first quarter of 2016.

In 2013, the ARJEL drafted a report including 33 proposals to fight against excessive online gambling [5]. One of them concerned the introduction of a system for identifying problematic and disordered gamblers based on indicators available in the ARJEL's database. The aim of this study is to contribute to this proposal by developing a model for the identification of excessive gamblers, and to extend it to all authorized gambling types.

Several studies have already been performed using a database from a single online gambling operator: data from Bwin were used by [6], [7], [8], [9] and [10] , data from GTECH G2 were used in [11] and data from Winamax were used in [12].

In most of them, one of the objectives was to identify potential problem gamblers and/or find indicators or behaviors associated to problem gambling. As the gamblers' status (e.g. non problem or problem gambler) were *a priori* unknown, two types of methodology had been considered. The first method, applied by [11] and [10], was to cluster the gamblers into homogenous classes. Then, by interpreting the classes, the authors could identify groups of gamblers with potentially risky behaviors. However, as highlighted in [10], "*it is difficult to assign any of the clusters to specific clinical groups with a high degree of certainty with respect to problem gambling*". The second solution was to approximate the gamblers' status by a proxy. In this case, the chosen proxy served as a reference to compare groups or predict the value of the proxy by using gambling data. For example, in [6], the authors compared the top 1% most involved gamblers (in terms of amount wagered or number of bets) to the remaining 99%. In [7] and [9], the authors compared the behavior of gamblers who closed their accounts because of gambling-related problems to the behavior of other account closers. The relevance of these different proxies is questionable, and, as stated by [10], a solution could be "*integrating gambling data with psychological studies using structured interviews in order to determine the level of gambling problems*".

In a recent study using the Winamax database [12], the authors collected actual gambling data from a sample of poker players as well as responses to the Problem Gambling Severity Index (PGSI) [13]. Gamblers were separated in a non-problem gamblers group (PGSI $< 5$) and a problem gamblers group (PGSI $\geq 5$). A multivariate logistic regression was performed on gambling data to estimate the probability that an individual was screened as a problem gambler. However, even if the PGSI is often considered as a good screening tool, the clinical diagnosis based on the DSM-5 criteria remains the gold standard to assess gambling disorders.

With regard to the results of this study, the specificity of the model developed was only 49.3 %. Although the authors argue that 75 % of false positive gamblers had responded positively to at least one question on the PGSI, a more discriminant model might have been achieved by using another type of algorithm. In particular, the authors explain that quantitative variables were categorized into quartiles, which lead to a loss of information, among other problems. Furthermore, in a prediction perspective, it is often worth to train multiple models and select the best one. Moreover, like most of the studies using real gambling data, this work focused on only one type of gambling from a single operator, whose clients may not be representative of all online gamblers in a country.

In conclusion, several studies have been performed previously to identify problem online gambling, but they displayed multiple limits. The EDEIN study aims at overcoming these weaknesses by proposing a model for an early prediction of online gambling problems based on real-life gambling data, applicable to all types of gambling and clinically validated.

## METHODS AND ANALYSIS

The general objective of the study is to develop a model for screening excessive gambling practices based on the behavior of online gamblers. The research is split into three successive stages.

### First stage

The first stage aims at defining a typology of French online gamblers based on their real-life gambling data. The objective is to identify classes of gamblers with gambling behaviors that are potentially indicative of excessive gambling.

The study population will be a sample of 20,000 users of French gambling websites certified by the ARJEL (n=10 000) and from the FDJ (n=10 000). Only validated accounts for which the gambler has placed at least one bet during the inclusion period will be included. The set of available variables will include socio-demographic data (age and sex), data concerning the accounts (e.g. modification of bet limits, number of money deposits) and data related to gambling itself (e.g. number of active betting days, amount of wagers, net loss). For several indicators, we will also measure their evolution between the three previous months and the last 30 days. This will allow us to assess intra-gambler variability and detect individuals who changed their gambling habits early, in particular changes in risk taking (e.g. increase in net loss or number of money deposits). We will use these indicators to define a typology of online gamblers. Individuals with similar gambling practices will be grouped in the same class. By interpreting the classes, we might be able to find groups of presumably risky gamblers and identify the indicators most related to potential gambling problems.

The classification process will be based on the monthly data of gamblers. The statistical unit will be the data of a gambler observed over the last 30 days, and for certain indicators, the comparison between the month concerned and the three previous months. Accordingly, there will be 9 observations for each gambler, completed by the 3 previous months for each observation for evolutive-indicators. The same gambler will not necessarily be assigned to the same class for the 9 months. This period of 30 days allows obtaining enough data to estimate the gambling activity of a gambler correctly.

In order to take into account the specificity of each major type of gambling, five clustering models will be developed. Five sets of variables will be therefore taken into account (Table 1): set 1 is composed of variables “independent” of the type of game (e.g. money deposit, gambling limits), set 2 consists of variables which can be computed for each of the four types of games and which can be used either for the global gambling practice by summing all types of games or for each game separately (e.g. money wagered, use of bonus), set 3 is comprised of variables specific to sports betting (e.g. live betting, complex sports bets), set 4 contains variables relative to horse race betting (e.g. complex horse race bets) and set 5 is composed of variables specific to lotteries and scratch games (e.g. differed lottery). There were no variables specific to poker.

Table 1: Sets of variables used to define the five models

| | **Set 1 (variables « independent » from the type of game)** | **Set 2: (variables computed for each type of game)** | **Set 3 (variables specific to sports betting)** | **Set 4 (variables specific to horse race betting)** | **Set 5 (variables specific to lotteries and scratch games)** |
|---|---|---|---|---|---|
| | • Total deposit<br>• Total number of deposits<br>• Biggest deposit in a single day<br>• Total withdrawal<br>• Number of sequences of 3 deposits within a 12h period (chasing 1)<br>• Number of times when we observe a deposit made less than 1h after a bet (chasing 2)<br>• Number of different types of games played<br>• Number of changes in the wagering limits<br>• Number of changes in the deposit limits<br>• Number of changes in auto-withdrawal limits<br>• Highest limit set in wagering limits<br>• Highest limit set in deposit limits<br>• Number of active accounts | • Total stake<br>• Total number of bets<br>• Net loss<br>• Total win in the previous month<br>• Number of gambling days<br>• Coefficient of variation of the number of bets for gambling days<br>• Coefficient of variation of stakes for gambling days<br>• Biggest total stake in a single day<br>• Total bonuses used | • Total stake for complex bets<br>• Total stake for live-bets<br>• Number of different sports | • Total stake for complex bets | • number of different games played<br>• type of lottery (instant/deferred) |
| Global status | ✓ | ✓ (for all types of game) | | | |
| Sports betting | ✓ | ✓ (for sports betting) | ✓ | | |
| Horse race betting | ✓ | ✓ (for horse race betting) | | ✓ | |
| Poker | ✓ | ✓ (for poker) | | | |
| Lotteries and scratch games | ✓ | ✓ (for lotteries and scratch games) | | | ✓ |

The five sets of variables will be used to define five models (Table 1). The first model ("global status") will be based on variables of set 1 and variables of set 2 computed for the four types of gambling together; it represents the global activity of the gambler, irrespective of the gambling type. The second model ("sports betting") will be based on variables from set 1, variables from set 2 computed for sports betting and variables from set 3; it represents the gambling activity for sports betting only. The third model ("horse race betting") will use variables from set 1, variables from set 2 computed for horse race betting and indicators from set 4; it represents the gambling activity for horse race betting only. The fourth model will be based on variables from set 1 and variables from set 2 computed for poker; it represents the gambling activity for poker only. The fifth model will use variables from set 1, variables from set 2 for lotteries and scratch cards and variables from set 5; it represents the gambling activity for lotteries and scratch cards games only. As a result, we will obtain five classifications. For example, if a gambler who plays multiple games is in the cluster composed of potential problem gamblers in the global classification, it will be interesting to look at which cluster he/she belongs in the game-specific classifications. This strategy could lead to more preventive actions targeted toward the problematic games only.

The statistical method used for the classification of gamblers will be a latent class clustering analysis [14]. In this model-based clustering approach, we assume that the data is generated from a mixture of underlying probability distributions. In other words, we assume that the population of gamblers observed is heterogeneous, composed from several but *a priori* unknown homogenous sub-populations (e.g. casual gamblers, risky gamblers, etc.). One of the advantages of latent class clustering is the possibility to compute statistical criteria based on the likelihood of the model (AIC, BIC) or on the accuracy of the classification (entropy) to choose the number of clusters and evaluate the fit of the models. This approach also allows comparing different model specifications (for the same number of classes) by adding or relaxing some constraints as described in [15]. These constraints are for instance considering that the variance of (continuous) observed variables are invariant in the different classes or assuming that observed variables are independent in each cluster (local independence). Adding such constraints allows reducing the number of parameters of a latent class model which can grow rapidly with the number of classes. Moreover, latent class models allow estimating two types of probabilities: the probability to observe an individual's characteristic given the latent class and the probability that an individual belongs to a latent class given his/her characteristics. For example, we can estimate the probability that a gambler in class *k* makes 2 deposits of money in the month and the probability that a gambler who makes 2 deposits of money belongs to the class *k*. Another advantage of latent class clustering is the possibility to include both continuous and categorical variables via the use of appropriate distribution (typically Gaussian distributions for continuous indicators and binomial/multinomial distributions for binary/ordinal indicators).

After estimating the model parameters, groups of gamblers will be interpreted by describing the variables in each class by their mean, median or category probabilities according to the nature of the variable. Differences between clusters for each indicator will be assessed by global and pairwise tests.

**Second stage**

The objective of the second stage is to define a score to predict whether a gambler is problematic or not. The score will be obtained by crossing the gambling real-life data with the results of the PGSI, obtained from an online questionnaire.

The PGSI will be diffused to a panel of active online gamblers who agree to reply voluntarily and anonymously. Every gambler will therefore be assigned a status depending on the results of the PGSI (high risk of gambling problems, moderate risk, low risk or no risk). The gambling data of the respondents will be linked to the answers to the questionnaire thanks to the use of an encrypted identifier.

We expect about 16,000 gamblers to answer the questionnaire (10,000 for the ARJEL and 6,000 for the FDJ). Based on their score on the PGSI, gamblers will be classified in 1 of the 4 following categories: score of 0: no gambling problem; score of 1 or 2: low risk of gambling problem; score of 3 to 7: moderate risk of gambling problem; score of 8 or more: excessive gambling problem.

In order to predict gamblers' status defined by the PGSI, several supervised learning algorithms could be applied: for example, logistic regression, support vector machines (SVM) [16] and random forest [17]. A recent article [9] compared the performance of several data mining procedures to identify high-risk online sports gamblers (individuals who closed their account due to gambling-related problems). In particular, logistic regression, LASSO regression [18], artificial neural networks (ANN) [19], SVM and random forests were compared in terms of sensitivity, specificity, accuracy, precision and area under the curve (AUC). The results show that none of the algorithms tested gave acceptable results (the sensitivity ranged from to 1.8% to 29.1%). ANN was the method who performed the best in the hold-out sample with a sensitivity of 29.1% and a specificity of 81.1%. In addition to the comparison of statistical methods, one of the conclusions was that variables included in the model (i.e. variables from the Bwin database used in [20] and demographic variables) were insufficient to correctly predict problem gamblers. The author highlighted the need to identify new behavioral variables to build more efficient models. In our project, we attempt to overcome these weaknesses by integrating such behavioral variables (for example, chasing proxies or behavior changes over time), and by testing several supervised learning algorithms in order to select the most efficient one.

We will measure the sensitivity (probability to detect gamblers at risk), specificity (probability to detect gamblers not at risk), positive predictive value (probability that an individual detected at risk is really at risk), negative predictive value (probability that a gambler detected not at risk is really not at risk) and AUC for each method used. Other measurements, such as Brier's scores and calibration assessment, could be computed depending on the algorithm used. In order to limit overfitting, data will be partitioned into learning sample, validation sample and test sample or/and cross-validation will be applied, depending on the method used.

### Third stage

The third stage of the study aims at clinically validating the screening model obtained at the second stage, by comparing the predictions of the model to current diagnosis of gambling disorder based on the National Opinion Research Center DSM-IV Screen for Gambling Problems (NODS). We will use a version of the NODS that takes into account the changes in the gambling disorders section in the DSM-5. Furthermore, we will diagnose lifetime and past 12 months gambling disorders. Patients with current gambling problems will be asked about the presence of symptoms for each of the past 12 months.

Participation in stage 3 will be proposed after completing the online questionnaire at stage 2. Once eligibility is confirmed, and if the volunteer actually accepts to take part in stage 3, a telephone clinical interview will be offered, lasting about 30 minutes. It will allow defining the gambler's clinical status (presence or absence of a gambling disorder diagnosis), according to the DSM-5 definition. The interview will also comprise a set of questions on the gambling course and habits, motivations, gambling-related cognition, and negative consequences.

We expect at least 240 participants (60 for each type of gambling) including half with a gambling problem detected by the PGSI (threshold of 8) and half with no gambling problem.

The diagnosis obtained from the telephone clinical interview will be crossed with the screening model results to assess the clinical validity of the model. As for stage 1 and 2, analyses will be carried out separately between the principal 4 types of gambling: poker, sports betting, horse race betting, and lotteries and scratch games. This third stage will also eventually allow identifying new or improved proxies of risky gambling behavior, especially based on time indicators (which are not available in the data sets due to difficulties to extract them; for example: duration of gambling per day, chasing, etc.). A summary of data collected through the three stages of the project is shown in Table 2.

Table 2: Summary of data collected throughout the three phases

| Phase 1 | Phase 2 | Phase 3 |
|---|---|---|
| • Gambling data from the ARJEL's and the FDJ's database (i.e. variables from Table 1) | • Data obtained from the online questionnaire<br>• Gambling data of participants to phase 2 (new data extraction by the ARJEL and the FDJ) | • Sociodemographic data<br>• Variables on gambling habits<br>• Evaluation of cognitive distortions<br>• Diagnosis of gambling disorders<br>• Gambling data obtained from participants' account history (especially time-related variables)<br>• Gambling data of participants to phase 3 (new data extraction by the ARJEL and the FDJ) |

## DISCUSSION

The main objective of this study is to develop a model for screening problematic online gambling behaviors using player account data. Gambling disorders will be assessed by the PGSI and results will be crossed with gamblers data and validated clinically with DSM-5-based diagnoses. Latent class clustering and supervised learning algorithms will be used to estimate several models and choose the best one.

Since all online gambling operators authorized in France will be involved, data will be fully representative of the French gamblers' population, overcoming the limitations of previous studies based on data from a single gambling operator. This will increase the generalizability of the results and enable routine use of the screening model on any gambling website. Another strength of this project is the clinical validation of the screening model assessed through a clinical interview of a subset of gamblers. This is the first time that a diagnosis-based clinical validity (rather than a screening with an auto-questionnaire) is used for this kind of study, thus bringing to the model developed a high construct validity. This study also takes into account the specificity of gambling types which allows detecting gambling problems at an individual level (with the overall gambling practice, whatever the gambling type) and separately for the different types of online gambling (poker, sports betting, horse race betting and lotteries and scratch cards).

Moreover, variables used in stages 1 and 2 are not just raw data but rather have been defined specifically to be related to risky gambling behaviors.

In the first stage, latent class models may produce different results compared to more traditional methods, like k-means or hierarchical clustering, used in the context of online gambling data. In particular, these latter methods sometimes generate several very homogenous clusters and a cluster grouping all extreme profiles. However, problem gambling behavior is considered as one particular extreme behavior that we want to identify precisely. The use of latent variable models can help to detect this kind of behavior. In the second stage, the use of different kinds of supervised algorithms may allow identifying the best performing model(s) to predict individuals with gambling problems. Moreover, it will be interesting to compare the clusters obtained from the latent class analysis with the prediction of the screening model.

Some limitations of this study include potential selection bias at stages 2 and 3. To uncover these biases, real-life gambling data will be compared between individuals who completed the PGSI (stage 2) and a representative sample of non-selected gamblers (stage 1), and between gamblers interviewed (stage 3) and those not interviewed during the clinical phase (stages 1 and 2). Furthermore, real-life gambling data will not include time-based indicators, because of the complexity of their calculation on a large scale. This limit have been highlighted in previous studies as a potential bias [12]. Moreover, building new behavioral variables was recommended to develop more efficient models [9]. By having access to the gamblers' account history during the clinical interviews, we will collect more detailed data (especially time-related data) and may be able to find new proxies of risky gambling behaviors, such as chasing behavior or time spent gambling.

## CONCLUSION

Thus, the results of this study may lead to the implementation of a system for the early identification of problem and disordered gamblers based on real-life gambling data. The project will therefore contribute to setting up an innovative and effective prevention measure, to inform and protect gamblers as early as possible. This tool will have the advantage of being usable in quasi-routine on any gambling website since no intervention from the gamblers will be required to define their status. Specific information and advice could be early provided for individuals identified as at risk or problem gamblers.

## LIST OF ABBREVIATIONS

DSM-5: Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition; ARJEL: Authority for Regulating Online Gambling; FDJ: Française Des Jeux; PGSI: Problem Gambling Severity Index; SVM: support vector machine; ANN: artificial neural network; AUC: area under the curve

## DECLARATIONS

### Ethics approval and consent to participate

For stage 2 and 3, participants were informed about the research. For stage 3, they gave their written informed consent prior to their inclusion in the study. This study was approved by the local Research Ethics Committee (GNEDS) on March 25, 2015.

### Dissemination

Results will be presented in national and international conferences, submitted to peer-reviewed journals and will be part of a PhD thesis. A final report with the study results will be presented to the ARJEL, especially the final screening model. This final report will be written in reference to the ARJEL's report to combat excessive online gambling.

### Competing interests

MGB, JC and GCB declare that the University Hospital of Nantes has received funding from gambling industry (FDJ and PMU) in the form of a sponsorship which supports the gambling section of the BALANCED Unit (the Reference Centre for Excessive Gambling). Scientific independence towards gambling industry operators is warranted. There were no constraints on publishing.

BP, JMC and JBH declare that they have no conflict of interest.

## Funding

This research has benefited from the joint assistance of the French National Health Insurance Fund for Employees (CNAMTS), the French Directorate General of Health (DGS), the Arc Foundation for Cancer Research, the French National Cancer Institute (INCA), the French National Institute for Prevention and Education in Health (INPES), the French National Institute of Health and Medical Research (INSERM), the French Inter-Departmental Agency for the Fight against Drugs and Addictive Behaviors (Mildeca) and the French Social Security Scheme for Liberal Professionals (RSI) as part of the "Primary Prevention" call for proposals issued by the French Institute for Public Health Research (IReSP) and INCA in 2013.

## Authors' contributions

BP conducted the literature research on statistical methods and will conduct the statistical analysis.

JBH provided methodological advice, designed the statistical analysis plan and will supervise the statistical analysis.

JMC and GCB selected the parameters to be included in the study.

JMC designed and was responsible of the questionnaire diffusion in stage 2 to ARJEL gamblers.

JC gave advices about Internet gambling and prevention.

MGB is the principal investigator and provide medical supervision of the study.

GCB designed the study, wrote the protocol and is responsible for the project management.

BP wrote the first draft of the manuscript and all authors read and approved the final manuscript.

## Acknowledgements

We would like to warmly thank the ARJEL and the FDJ for their contribution to this project, by allowing us to have access to the gamblers' data.

This research was conducted at the initiative of and coordinated by the Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders" of the University Hospital of Nantes, the sponsor of this study.

## REFERENCES

1 Association AP. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®).* American Psychiatric Pub 2013.

2 Costes J-M, Eroukmanoff V, Richard J-B, *et al.* Les jeux d'argent et de hasard en France en 2014. Note de l'ODJ n°6. 2015.

3 Tovar ML, Costes JM, Eroukmanoff V. Les jeux d'argent et de hasard sur internet en France en 2012. *Tendances* 2013;**85**:1–6.

4 Griffiths M. Internet gambling: issues, concerns, and recommendations. *Cyberpsychol Behav* 2003;**6**:557–68.

5 ARJEL. Lutter contre le jeu excessif ou pathologique. Recommandations trois ans après l'adoption de la loi d'ouverture du marché des jeux en ligne. http://www.arjel.fr/IMG/pdf/20130426-addiction.pdf

6 LaBrie RA, LaPlante DA, Nelson SE, *et al.* Assessing the Playing Field: A Prospective Longitudinal Study of Internet Sports Gambling Behavior. *J Gambl Stud* 2007;**23**:347–62. doi:10.1007/s10899-007-9067-3

7 LaBrie R, Shaffer HJ. Identifying behavioral markers of disordered Internet sports gambling. *Addict Res Theory* 2011;**19**:56–65. doi:10.3109/16066359.2010.512106

8 Broda A, LaPlante DA, Nelson SE, *et al.* Virtual harm reduction efforts for Internet gambling: effects of deposit limits on actual Internet sports gambling behavior. *Harm Reduct J* 2008;**5**:27. doi:10.1186/1477-7517-5-27

9 Philander KS. Identifying high-risk online gamblers: a comparison of data mining procedures. *Int Gambl Stud* 2014;**14**:53–63. doi:10.1080/14459795.2013.841721

10 Adami N, Benini S, Boschetti A, *et al.* Markers of unsustainable gambling for early detection of at-risk online gamblers. *Int Gambl Stud* 2013;**13**:188–204. doi:10.1080/14459795.2012.754919

11 Dragicevic S, Tsogas G, Kudic A. Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection. *Int Gambl Stud* 2011;**11**:377–91. doi:10.1080/14459795.2011.629204

12 Luquiens A, Tanguy M-L, Benyamina A, *et al.* Tracking online poker problem gamblers with player account-based gambling data only. *Int J Methods Psychiatr Res* 2016;:n/a-n/a. doi:10.1002/mpr.1510

13 Ferris J, Wynne H. The Canadian problem gambling index. *Ott Can Cent Subst Abuse* 2001.

14 Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, eds. Applied Latent Class Analysis. Cambridge: Cambridge University Press;2002;89–106.

15 Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* 1998;**41**:578–588.

16 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.

17 Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.

18 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 1996;:267–88.

19 Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958;**65**:386.

20 Braverman J, Shaffer HJ. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. *Eur J Public Health* 2010;:ckp232. doi:10.1093/eurpub/ckp232

BMJ Open

# Study protocol for a transversal study to develop a screening model for excessive gambling behaviors on a representative sample of users of French authorized gambling websites.

| | |
|---|---|
| Journal: | *BMJ Open* |
| Manuscript ID | bmjopen-2016-014600.R1 |
| Article Type: | Protocol |
| Date Submitted by the Author: | 14-Feb-2017 |
| Complete List of Authors: | Perrot, Bastien; University of Nantes, EA 4275 SPHERE<br>Hardouin, Jean-Benoit; University of Nantes, EA 4275 SPHERE<br>Costes, Jean-Michel; Observatoire Des Jeux<br>Caillon, Julie; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders"<br>Grall-Bronnec, Marie; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders"<br>Challet-Bouju, Gaëlle; University Hospital of Nantes, Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders" |
| <b>Primary Subject Heading</b>: | Addiction |
| Secondary Subject Heading: | Public health |
| Keywords: | Gambling, Prevention, Latent class model, PUBLIC HEALTH, Online gambling |
| | |

# Study protocol for a transversal study to develop a screening model for excessive gambling behaviors on a representative sample of users of French authorized gambling websites.

Bastien PERROT[1*], Jean-Benoit HARDOUIN[1], Jean-Michel COSTES[2], Julie CAILLON[3,1], Marie GRALL-BRONNEC[3,1], Gaëlle CHALLET-BOUJU[3,1]

[*] Correspondence:
Bastien Perrot
bastien.perrot@univ-nantes.fr
EA 4275 - IRS2, 22 boulevard Bénoni Goullin – 44200 Nantes, France
Tel. +33 253009129

[1] EA 4275 SPHERE "methodS in Patient-centered outcomes and HEalth ResEarch", Faculties of Medicine and Pharmaceutical Sciences, University of Nantes, France

[2] Observatoire des jeux, Ministère des Finances, Paris, France

[3] Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders", Department of Addictology and Psychiatry, University Hospital of Nantes, France

## ABSTRACT

### Introduction

Since the legalization of online gambling in France in 2010, gambling operators must implement responsible gambling measures to prevent excessive gambling practices. However, actually there is no screening procedure for identifying problematic gamblers. Although several studies have already been performed using several datasets from online gambling operators, the authors deplored several methodological and clinical limits that prevent scientifically validating the existence of problematic gambling behavior. The aim of this study is to develop a model for screening excessive gambling practices based on the gambling behaviors observed on French gambling websites, coupled with a clinical validation.

### Methods and analysis

The research is divided into three successive stages. All analyses will be performed for each major type of authorized online gambling in France. The first stage aims at defining a typology of users of French authorized gambling websites based on their gambling behavior. This analysis will be based on data from the Authority for Regulating Online Gambling (ARJEL) and the Française Des Jeux (FDJ). For the second stage aiming at determining a score to predict whether a gambler is problematic or not, we will cross answers from the Canadian Problem Gambling Index with real gambling data. The objective of the third stage is to clinically validate the score previously developed. Results from the screening model will be compared (using sensitivity, specificity, AUC, and positive and negative predictive values) to the diagnosis obtained with a telephone clinical interview, including diagnostic criteria for gambling addiction.

### Ethics and dissemination

This study was approved by the local Research Ethics Committee (GNEDS) on March 25, 2015. Results will be presented in national and international conferences, submitted to peer-reviewed journals and will be part of a PhD thesis. A final report with the study results will be presented to the ARJEL, especially the final screening model.

## Strengths and limitations of this study

- This study will contribute to setting up an innovative prevention measure used to inform and protect gamblers as early as possible.
- Data will be representative of the French online gamblers' population; all types of authorized online gambling forms, and data from all authorized operators in France will be included.
- The final screening model will be clinically validated.
- We will use advanced statistical methods to build the screening model
- Selection bias may occur during participants' recruitment at stages 2 and 3.
- Due to technical and confidentiality constraints, some potentially interesting gambling indicators will not be used.

## Trial registration

NCT02415296

**Keywords**

Online gambling, problem gambling, prevention, classification, latent class model, predictive model.

## INTRODUCTION

The majority of gamblers have a controlled and recreational gambling practice, but some of them lose control of it. In the Fifth Edition of the Diagnostic and Statistical Manual of Mental Disorders, (DSM-5), gambling disorder is defined as a "persistent and recurrent problematic gambling behavior leading to clinically significant impairment or distress" [1]. A French national survey conducted in 2015 indicated that 3.9% of past-year gamblers had gambling problems, including 0.9% of excessive gamblers [2]. With regard to online gambling only, Tovar et al. [3] found a prevalence of gambling problems of 17% among Internet gamblers, including 6.6% of excessive gamblers. Similar rates were found by Wood and Williams [4] among a sample of Canadian gamblers; indeed, the authors of this study found that 5.7% of non-Internet gamblers could be considered as problem gamblers, including 1.7% of excessive gamblers, compared to a prevalence of gambling problems of 16.6% among Internet gamblers, including 3.8% of excessive gamblers. Several factors may explain the fact that online gambling is possibly more addictive than offline gambling; especially greater accessibility, increased disinhibition and higher event frequency [5].

The opening to competition and regulation of the French online gambling sector was provided by the law No 2010-476 of 12 May 2010. This law included the creation of the Regulatory Authority for Online Gambling (ARJEL in French). The ARJEL can issue licenses to online operators for only three types of games: horse race betting, sports betting and poker. Lotteries and scratch games provided by the Française des Jeux (FDJ, i.e. France's national lottery operator) are also authorized online because of a particular waiver. The ARJEL compiles data from all accounts created on authorized online gambling sites in France, namely 1,352,000 active accounts in the first quarter of 2016.

In 2013, the ARJEL drafted a report including 33 proposals to fight against excessive online gambling [6]. One of them concerned the introduction of a system for identifying problematic and disordered gamblers based on indicators available in the ARJEL's database. The aim of this study is to contribute to this proposal by developing a model for the identification of excessive gamblers and extend it to all authorized gambling types.

Several studies have already been performed using a database from a single online gambling operator: data from Bwin were used by LaBrie et al. [7], LaBrie & Shaffer [8], Broda et al. [9], Philander [10] and Adami et al. [11], data from GTECH G2 were used by Dragicevic et al. [12] and data from Winamax were used by Luquiens et al. [13].

In most of them, one of the objectives was to identify potential problem gamblers and/or find indicators or behaviors associated to problem gambling. As the gamblers' status (e.g. non-problem or problem gambler) were *a priori* unknown, two types of methodology had been considered. The first method, applied by Dragicevic et al. [12] and Adami et al. [11], was to cluster the gamblers into homogenous classes. Then, by interpreting the classes, the authors could identify groups of gamblers with potentially risky behaviors. However, as highlighted by Adami et al. [11], "*it is difficult to assign any of the clusters to specific clinical groups with a high degree of certainty with respect to problem gambling*". The second solution was to approximate the gamblers' status by a proxy. In this case, the chosen proxy served as a reference to compare groups or predict the value of the proxy by using gambling data. For example, LaBrie et al. [7],  compared the top 1% most involved gamblers (in terms of amount wagered or number of bets) to the remaining 99%. LaBrie et al. [8] and Philander [10] compared the behavior of gamblers who closed their accounts because of gambling-related problems to the behavior of other account closers. The relevance of these various proxies is questionable, and, as stated by [11], a solution could be "*integrating gambling data with* *psychological studies using structured interviews in order to determine the level of gambling problems*".

In a recent study using the Winamax database [13], the authors collected actual gambling data from a sample of poker players as well as responses to the Problem Gambling Severity Index (PGSI) [14]. Gamblers were separated into a non-problem gamblers group (PGSI < 5) and a problem gamblers group (PGSI ≥ 5). A multivariate logistic regression was performed on gambling data to estimate the probability that an individual was screened as a problem gambler. However, even if the PGSI is often considered as a good screening tool, the clinical diagnosis based on the DSM-5 criteria remains the gold standard in assessing gambling disorders.

With regard to the results of this study, the specificity of the model developed was only 49.3 %. Although the authors argue that 75 % of "false positive gamblers" (i.e. the gamblers wrongly classified as problem gamblers, 50.7%) had responded positively to at least one question on the PGSI, a more discriminant model might have been achieved by using another type of algorithm. In particular, the authors explain that quantitative variables were categorized into quartiles, which lead to a loss of information compared to treating them as continuous. Furthermore, in a prediction perspective, it is often worth to train multiple models and select the best one. Moreover, like most of the studies using real gambling data, this work focused on only one type of gambling from a single operator, whose clients may not be representative of all online gamblers in a given country.

In conclusion, several studies have been previously performed to identify problem online gambling, but they displayed multiple limits. The EDEIN (Screening for Excessive Gambling Behaviors on the Internet) study aims at overcoming these weaknesses by proposing a model for an early prediction of online gambling problems based on player account-based gambling data, applicable to all types of gambling and clinically validated. Our objective is also to identify distinct gamblers' profiles based on their gambling behavior in order to set up targeted prevention measures.

## METHODS AND ANALYSIS

The general objective of the study is to develop a model for screening excessive gambling practices based on the behavior of online gamblers. The research is split into three successive stages.

### First stage

The first stage aims at defining a typology of French online gamblers based on their player account-based gambling data. The objective is to identify classes of gamblers with gambling behaviors that are potentially indicative of excessive gambling.

The study population will be an anonymized random sample of 20,000 users of French gambling websites certified by the ARJEL (n=10,000) and from the FDJ (n=10,000). Only validated accounts for which the gambler has placed at least one bet during the inclusion period will be included. The set of available variables will include socio-demographic data (age and sex), data concerning the accounts (e.g. modification of bet limits, number of money deposits) and data related to gambling itself (e.g. number of active betting days, amount of wagers, net loss). Despite their potential interest for screening excessive gambling practice, some indicators cannot be extracted due to technical (e.g. poker multitabling, duration of sessions, chasing) or confidentiality constraints (e.g. amount of each deposit and bet). For several indicators, we will also measure their evolution between the three previous months and the last 30 days. This will allow us to assess intra-gambler variability and detect individuals who changed their gambling habits early, especially risk taking changes (e.g. increase in net loss or number of money deposits). We will use these indicators to define a typology of online gamblers. Individuals with similar gambling practices (e.g. similar gambling frequencies, similar number of deposits per month, etc.) will be grouped in the same class. By interpreting the classes, we might be able to find groups of presumably risky gamblers and identify the indicators most related to potential gambling problems.

The classification process will be based on the monthly data of gamblers. The statistical unit will be the data of a gambler observed over the last 30 days and, for certain indicators (e.g. total stake, total number of bets, total deposit, etc.), a comparison between the given month and the three previous months. Accordingly, to classify a gambler at month *m*, we will use his gambling activity during month *m* plus the evolution (expressed as a difference or a ratio according to the indicators) of gambling activity between *m* and the average gambling activity measured at *m-3*, *m-2* and *m-1*. Thus, as we need a 3-month "follow-up" for each monthly classification, gamblers will not be classified for the first 3 months of the inclusion period. . The same gambler will not necessarily be assigned to the same class for the 9 months.

In order to take into account the specificity of each major type of gambling, five clustering models will be developed. Five sets of variables will therefore be taken into account (Table 1): set 1 is composed of variables that are "independent" of the type of game (e.g. money deposit, gambling limits), set 2 consists of variables which can be computed for each of the four types of games and can be used either for the global gambling practice by summing all types of games or for each game separately (e.g. money wagered, use of bonus), set 3 is comprised of specific sports betting variables (e.g. live betting, complex sports bets), set 4 contains horse race betting variables (e.g. complex horse race bets) and set 5 is composed of specific lottery and scratch game variables (e.g. differed lottery). There were no specific poker variables available.

Table 1: Sets of variables used to define the five models

| | **Set 1 (variables "independent" from the type of game)** | **Set 2: (variables computed for each type of game)** | **Set 3 (specific sports betting variables)** | **Set 4 (specific horse race betting variables)** | **Set 5 (variables specific to lotteries and scratch games)** |
|---|---|---|---|---|---|
| | • Total deposit<br>• Total number of deposits<br>• Biggest deposit in a single day<br>• Total withdrawal<br>• Number of sequences of 3 deposits within a 12h period (chasing 1)<br>• Number of times when we observe a deposit made less than 1h after a bet (chasing 2)<br>• Number of different types of games played<br>• Number of changes in the wagering limits<br>• Number of changes in the deposit limits<br>• Number of changes in auto-withdrawal limits<br>• Highest limit set in wagering limits<br>• Highest limit set in deposit limits<br>• Number of active accounts | • Total stake<br>• Total number of bets<br>• Net loss<br>• Total win in the previous month<br>• Number of gambling days<br>• Coefficient of variation of the number of bets for gambling days<br>• Coefficient of variation of stakes for gambling days<br>• Biggest total stake in a single day<br>• Total bonuses used | • Total stake for complex bets<br>• Total stake for live bets<br>• Number of different sports | • Total stake for complex bets | • number of different games played<br>• type of lottery (instant/deferred) |
| Global status | ✓ | ✓ (for all types of game) | | | |
| Sports betting | ✓ | ✓ (for sports betting) | ✓ | | |
| Horse race betting | ✓ | ✓ (for horse race betting) | | ✓ | |
| Poker | ✓ | ✓ (for poker) | | | |
| Lotteries and scratch games | ✓ | ✓ (for lotteries and scratch games) | | | ✓ |

The five sets of variables will be used to define five models (Table 1). The first model ("global status") will be based on variables of set 1 and variables of set 2 computed for the four types of gambling together; it represents the global activity of the gambler, irrespective of the gambling type. The second model ("sports betting") will be based on variables from set 1, variables from set 2 computed for sports betting and variables from set 3; it represents the gambling activity for sports betting only. The third model ("horse race betting") will use variables from set 1, variables from set 2 computed for horse race betting and indicators from set 4; it represents the gambling activity for horse race betting only. The fourth model will be based on variables from set 1 and variables from set 2 computed for poker; it represents the gambling activity for poker only. The fifth model will use variables from set 1, variables from set 2 for lotteries and scratch cards and variables from set 5; it represents the gambling activity for lotteries and scratch card games only. As a result, we will obtain five classifications. For example, if a gambler who plays multiple games is in the cluster composed of potential problem gamblers in the global classification, it will be interesting to look at which cluster he/she belongs to in the game-specific classifications. This strategy could lead to more preventive actions targeted toward the problematic games only.

The statistical method used for the classification of gamblers will be a latent class clustering analysis [15]. In this model-based clustering approach, we assume that the data is generated from a mixture of underlying probability distributions. In other words, we assume that the population of gamblers observed is heterogeneous, composed from several but *a priori* unknown homogenous sub-populations (e.g. casual gamblers, risky gamblers, etc.). One of the advantages of latent class clustering is the possibility to compute statistical criteria based on the likelihood of the model (e.g. Akaike Information Criterion and Bayesian Information Criterion) or accuracy of the classification (entropy) in order to choose the number of clusters and evaluate the fit of the models. This approach also allows to compare various model specifications (for the same number of classes) by adding or relaxing some constraints as described in [16]. These constraints are, for instance, considering that the variance of (continuous) observed variables are invariant in the different classes or assuming that observed variables are independent in each cluster (local independence). Adding such constraints allows to reduce the number of parameters of a latent class model which can grow rapidly with the number of classes. Moreover, latent class models allow to estimate two types of probabilities: the probability to observe an individual's characteristic given the latent class and the probability that an individual belongs to a latent class given his/her characteristics. For example, we can estimate the probability that a gambler in class *k* makes 2 deposits of money in the month and the probability that a gambler who makes 2 deposits of money belongs to class *k*. Another advantage of latent class clustering is the possibility of including both continuous and categorical variables via the use of appropriate distribution (typically Gaussian distributions for continuous indicators and binomial/multinomial distributions for binary/ordinal indicators).

After estimating the model parameters, groups of gamblers will be interpreted by describing the variables in each class by their mean, median or category probabilities according to the nature of the variable. Differences between clusters for each indicator will be assessed by global and pairwise tests.

**Second stage**

The objective of the second stage is to define a score in order to predict whether a gambler is problematic or not. The score will be obtained by crossing the player account-based gambling data with the results of the PGSI, obtained from an online questionnaire.

The PGSI will be issued to a panel of active online gamblers who agree to reply voluntarily and anonymously. Every gambler will therefore be assigned a status depending on the results of the PGSI (high risk of gambling problems, moderate risk, low risk or no risk). The gambling data of the respondents will be linked to the answers to the questionnaire thanks to the use of an encrypted identifier.

Based on an estimated 1–3% response rate, we expect about 20,000 gamblers to answer the questionnaire (10,000 for the ARJEL and 10,000 for the FDJ). Based on their score on the PGSI, gamblers will be classified in 1 of the 4 following categories: score of 0: no gambling problem; score of 1 or 2: low risk of gambling problem; score of 3 to 7: moderate risk of gambling problem; score of 8 or more: excessive gambling problem.

In order to predict gamblers' status defined by the PGSI, several supervised learning algorithms could be applied: for example, logistic regression, support vector machines (SVM) [17] and random forest [18]. A recent article [10] compared the performance of several data mining procedures to identify high-risk online sports gamblers (individuals who closed their account due to gambling-related problems). In particular, logistic regression, LASSO regression [19], artificial neural networks (ANN) [20], SVM and random forests were compared in terms of sensitivity, specificity, accuracy, precision and area under the curve (AUC). The results show that none of the algorithms tested gave acceptable results (the sensitivity ranged from to 1.8% to 29.1%). ANN was the method who performed the best in the hold-out sample with a sensitivity of 29.1% and a specificity of 81.1%. In addition to the comparison of statistical methods, one of the conclusions was that variables included in the model (i.e. variables from the Bwin database used in [21] and demographic variables) were insufficient to correctly predict problem gamblers. The author highlighted the need to identify new behavioral variables to build more efficient models. In our project, we attempt to overcome these weaknesses by integrating such behavioral variables (for example, chasing proxies or behavior changes over time), and by testing several supervised learning algorithms in order to select the most efficient one.

We will measure the sensitivity (probability to detect gamblers at risk), specificity (probability to detect gamblers not at risk), positive predictive value (probability that an individual detected at risk is really at risk), negative predictive value (probability that a gambler detected not at risk is really not at risk) and AUC for each method used. Other measurements, such as Brier's scores and calibration assessment, could be computed depending on the algorithm used. In order to limit overfitting, data will be partitioned into learning sample, validation sample and test sample or/and cross-validation will be applied, depending on the method used.

**Third stage**

The third stage of the study aims at clinically validating the screening model obtained at the second stage, by comparing the predictions of the model to current diagnosis of gambling disorder based on the National Opinion Research Center DSM-IV Screen for Gambling Problems (NODS) [22]. We will use a revised version of the NODS that we have created to take into account the changes in the gambling disorders section in the DSM-5. Furthermore, we will diagnose lifetime and past 12 months gambling disorders. Individuals with current gambling problems will be asked about the presence of symptoms for each of the past 12 months.

Participation in stage 3 will be proposed after completing the online questionnaire at stage 2. We will also propose participation in stage 3 to gamblers registered in our clinical unit's volunteer base. Once eligibility is confirmed, and if the volunteer actually accepts to take part in stage 3, a telephone clinical interview conducted by well-trained staff members with experience with pathological gamblers will be offered, lasting about 30 minutes. It will make it possible to define the gambler's clinical status (presence or absence of a gambling disorder diagnosis), according to the DSM-5 definition. The interview will also comprise a set of questions on the gambling course and habits, motivations (assessed by the Gambling Motives Questionnaire-Financial [23]), gambling-related cognitions (assessed by the Gambling Related Cognitions Scale [24,25], and negative consequences (assessed by questions measuring the impact of gambling on different areas of life and a scale currently being validated). Participants will be given a 50 € gift voucher in compensation for their participation.

We expect at least 240 participants (60 for each type of gambling), including half with a gambling problem detected by the PGSI (threshold of 8) and half with no gambling problem. The total number of participants was computed so that we obtain at least 30 individuals by group. The diagnosis obtained from the telephone clinical interview will be crossed with the screening model results to assess the clinical validity of the model. As for stage 1 and 2, analyses will be carried out separately between the principal 4 types of gambling: poker, sports betting, horse race betting, lotteries and scratch games. This third stage will also eventually allow the identification of new or improved proxies of risky gambling behavior, especially based on time indicators (which are not available in the data sets due to difficulties to extract them; for example: duration of gambling per day, chasing, etc.). A summary of data collected through the three stages of the project is shown in Table 2.

Table 1: Summary of data collected throughout the three phases

| Phase 1 | Phase 2 | Phase 3 |
|---|---|---|
| • Gambling data from a random panel of gamblers' accounts extracted from the ARJEL and FDJ databases (i.e. variables from Table 1) | • Data obtained from the online questionnaire<br>• Gambling data of participants to phase 2 (new data from the ARJEL and FDJ databases) | • Sociodemographic data<br>• Variables on gambling habits<br>• Evaluation of cognitive distortions<br>• Diagnosis of gambling disorders<br>• Gambling data obtained from participants' account history (especially time-related variables)<br>• Gambling data of participants to phase 3 (new ARJEL and FDJ data extraction) |

## DISCUSSION

The main objective of this study is to develop a model for screening problematic online gambling behaviors using player account data. Gambling disorders will be assessed by the PGSI and results will be crossed with gamblers' data and validated clinically with DSM-5-based diagnoses. Latent class clustering and supervised learning algorithms will be used to estimate several models and choose the best one.

Since all online gambling operators authorized in France will be involved, data will be representative of the French gamblers' population, overcoming the limitations of previous studies based on data from a single gambling operator. This will increase the generalizability of the results and enable routine use of the screening model on any gambling website. Another strength of this project is the clinical validation of the screening model assessed through a clinical interview of a subset of gamblers. This is the first time that a diagnosis-based clinical validity (rather than screening via an auto-questionnaire) is used for this kind of study, thus bringing a high construct validity to the developed model. This study also takes into account the specificity of gambling types, which allows the detection of gambling problems at an individual level (with the overall gambling practice, whatever the gambling type) and separately for the different types of online gambling (poker, sports betting, horse race betting, lotteries and scratch cards).

Moreover, variables used in stages 1 and 2 are not just raw data but rather have been defined specifically to be related to risky gambling behaviors.

In the first stage, latent class models may produce different results compared to more traditional methods, like k-means or hierarchical clustering, used in the context of online gambling data. In particular, these latter methods sometimes generate several very homogenous clusters and a cluster grouping all extreme profiles. However, problem gambling behavior is considered as one particular extreme behavior that we want to identify precisely. The use of latent variable models can help to detect this kind of behavior. In the second stage, the use of different kinds of supervised algorithms may allow the identification of the best performing model(s) in order to predict individuals with gambling problems. Moreover, it will be interesting to compare the clusters obtained from the latent class analysis with the screening model prediction.

Some limitations of this study include potential self-selection bias [26] at stages 2 and 3. To uncover these biases, player account-based gambling data will be compared between individuals who completed the PGSI (stage 2) and a representative sample of non-selected gamblers (stage 1), as well as between gamblers interviewed (stage 3) and those not interviewed during the clinical phase (stages 1 and 2). With regard to the PGSI, the cutoffs for the low-risk and moderate-risk categories have been criticized in the literature and alternative cutoffs have been proposed to produce more discriminant intermediate categories [27]. It would be interesting to perform sensitivity analyses according to the definition of the two intermediate thresholds. Furthermore, the collected data will be representative of users of French authorized gambling websites but we will have no information about users of unlicensed gambling websites. A study by Costes et al. [28] showed that 46.3% of online gamblers gambled on at least one unlicensed site, of which 12.1% gambled exclusively on unlicensed sites. As another limitation, player account-based gambling data will not include time-based indicators, because of the complexity of their calculation on a large scale. This limit has been highlighted in previous studies as a potential bias [13]. Moreover, building new behavioral variables was recommended in order to develop more efficient models [10]. Thanks to the third clinical stage of the study, we will collect more detailed data (especially time-related data) and may be able to find new proxies of risky gambling behaviors, such as chasing behavior or time spent gambling, by gaining access to the gamblers' account history during clinical interviews.

## CONCLUSION

Thus, the results of this study may lead to the implementation of a system for the early identification of problem and disordered gamblers based on player account-based gambling data. The project will therefore contribute to setting up an innovative and effective prevention measure, in order to inform and protect gamblers as early as possible. In particular, the need for prevention tools has been highlighted for Internet gamblers [29], and the implementation of gambling moderators has been encouraged [30,31]. This tool will have the advantage of being quasi-routinely usable on any gambling website since no intervention from the gamblers will be required to define their status.

Specific information and advice could be provided early for individuals identified as at risk or problem gamblers. In addition to the screening of gambling problems, the identification of distinct gamblers' profiles based on their gambling behavior (e.g. individuals gambling a lot of money during short periods, gamblers with varying gambling behaviors, individuals playing many different games, etc.) will allow the implementation of targeted prevention measures. This includes messages focused on time spent gambling for gamblers with limited financial damage and high familial and social damage, information on the randomness of outcomes for gamblers with a high level of cognitive distortions or recommendation of pauses during the game for gamblers with an important chasing behavior. Moreover, a final report with the study results, especially the final screening model, will be presented to the ARJEL in reference to its report to combat excessive online gambling [6].

## LIST OF ABBREVIATIONS

DSM-5: Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition; ARJEL: Authority for Regulating Online Gambling; FDJ: Française Des Jeux; PGSI: Problem Gambling Severity Index; SVM: Support Vector Machine; ANN: Artificial Neural Network; AUC: Area Under the Curve

## DECLARATIONS

### Ethics approval and consent to participate

For stage 2 and 3, participants were informed about the research. For stage 3, they gave their written informed consent prior to their inclusion in the study. This study was approved by the local Research Ethics Committee (GNEDS) on March 25, 2015.

### Dissemination

Results will be presented in national and international conferences, submitted to peer-reviewed journals and will be part of a PhD thesis. A final report with the study results will be presented to the ARJEL, especially the final screening model. This final report will be written in reference to the ARJEL's report in order to combat excessive online gambling.

**Competing interests**

MGB, JC and GCB declare that the University Hospital of Nantes has received gambling industry (FDJ and PMU) funding in the form of a sponsorship which supports the gambling section of the BALANCED Unit (the Reference Centre for Excessive Gambling). Scientific independence towards gambling industry operators is warranted. There were no publishing constraints.

BP, JMC and JBH declare that they have no conflict of interest.

**Funding**

This research has benefited from the joint assistance of the French National Health Insurance Fund for Employees (CNAMTS), the French Directorate General of Health (DGS), the Arc Foundation for Cancer Research, the French National Cancer Institute (INCA), the French National Institute for Prevention and Education in Health (INPES), the French National Institute of Health and Medical Research (INSERM), the French Inter-Departmental Agency for the Fight against Drugs and Addictive Behaviors (Mildeca) and the French Social Security Scheme for Liberal Professionals (RSI) as part of the "Primary Prevention" call for proposals issued by the French Institute for Public Health Research (IReSP) and INCA in 2013.

**Authors' contributions**

BP conducted the literature research on statistical methods and will conduct the statistical analysis.

JBH provided methodological advice, designed the statistical analysis plan and will supervise the statistical analysis.

JMC and GCB selected the parameters to be included in the study.

JMC designed and was responsible of the questionnaire issued in stage 2 to ARJEL gamblers.

JC provided Internet gambling and prevention advice.

MGB is the principal investigator and provides the study's medical supervision.

GCB designed the study, wrote the protocol and is in charge of project management.

BP wrote the first draft of the manuscript and all authors read and approved the final manuscript.

## Acknowledgements

We would like to warmly thank the ARJEL and FDJ for their contribution to this project, by allowing us to have access to the gamblers' data.

This research is conducted on the initiative of and coordinated by the Clinical Investigation Unit BALANCED "BehaviorAL AddictioNs and ComplEx mood Disorders" of the University Hospital of Nantes, which sponsors this study.

## REFERENCES

1 Association AP. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub 2013.

2 Costes J-M, Eroukmanoff V, Richard J-B, *et al.* Les jeux d'argent et de hasard en France en 2014. *Notes ODJ N°6* 2015

3 Tovar ML, Costes JM, Eroukmanoff V. Les jeux d'argent et de hasard sur internet en France en 2012. *Tendances* 2013;**85**:1–6.

4 Wood RT, Williams RJ. A comparative profile of the Internet gambler: Demographic characteristics, game-play patterns, and problem gambling status. *New Media Soc* 2011;**13**:1123–41. doi:10.1177/1461444810397650

5 Griffiths M. Internet gambling: issues, concerns, and recommendations. *Cyberpsychol Behav* 2003;**6**:557–68.

6 ARJEL. Lutter contre le jeu excessif ou pathologique. Recommandations trois ans après l'adoption de la loi d'ouverture du marché des jeux en ligne. http://www.arjel.fr/IMG/pdf/20130426-addiction.pdf

7 LaBrie RA, LaPlante DA, Nelson SE, *et al.* Assessing the Playing Field: A Prospective Longitudinal Study of Internet Sports Gambling Behavior. *J Gambl Stud* 2007;**23**:347–62. doi:10.1007/s10899-007-9067-3

8 LaBrie R, Shaffer HJ. Identifying behavioral markers of disordered Internet sports gambling. *Addict Res Theory* 2011;**19**:56–65. doi:10.3109/16066359.2010.512106

9 Broda A, LaPlante DA, Nelson SE, *et al.* Virtual harm reduction efforts for Internet gambling: effects of deposit limits on actual Internet sports gambling behavior. *Harm Reduct J* 2008;**5**:27. doi:10.1186/1477-7517-5-27

10 Philander KS. Identifying high-risk online gamblers: a comparison of data mining procedures. *Int Gambl Stud* 2014;**14**:53–63. doi:10.1080/14459795.2013.841721

11 Adami N, Benini S, Boschetti A, *et al.* Markers of unsustainable gambling for early detection of at-risk online gamblers. *Int Gambl Stud* 2013;**13**:188–204. doi:10.1080/14459795.2012.754919

12 Dragicevic S, Tsogas G, Kudic A. Analysis of casino online gambling data in relation to behavioural risk markers for high-risk gambling and player protection. *Int Gambl Stud* 2011;**11**:377–91. doi:10.1080/14459795.2011.629204

13 Luquiens A, Tanguy M-L, Benyamina A, *et al.* Tracking online poker problem gamblers with player account-based gambling data only. *Int J Methods Psychiatr Res* 2016;:n/a-n/a. doi:10.1002/mpr.1510

14 Ferris J, Wynne H. The Canadian problem gambling index. *Ott Can Cent Subst Abuse* 2001.

15 Vermunt JK, Magidson J. Latent class cluster analysis. *Appl Latent Cl Anal* 2002;:89–106.

16 Fraley C, Raftery AE. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput J* 1998;**41**:578–588.

17 Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.

18 Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.

19 Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Methodol* 1996;:267–88.

20 Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958;**65**:386.

21 Braverman J, Shaffer HJ. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. *Eur J Public Health* 2010;:ckp232. doi:10.1093/eurpub/ckp232

22 Wickwire EM, Burke RS, Brown SA, *et al.* Psychometric Evaluation of the National Opinion Research Center DSM-IV Screen for Gambling Problems (NODS). *Am J Addict* 2008;**17**:392–5. doi:10.1080/10550490802268934

23 Devos G, Challet-Bouju G, Burnay J, *et al.* Adaptation and validation of the Gambling Motives Questionnaire-Financial (GMQ-F) in a sample of French-speaking gamblers. *Int Gambl Stud* 2016;**0**:1–15. doi:10.1080/14459795.2016.1264080

24 Raylu N, Oei TP. The Gambling Related Cognitions Scale (GRCS): Development, confirmatory factor validation and psychometric properties. *Addiction* 2004;**99**:757–69.

25 Grall-Bronnec M, Bouju G, Sebille-Rivain V, *et al.* A French adaptation of the Gambling-Related Cognitions Scale (GRCS): a useful tool for assessment of irrational thoughts among gamblers. *J Gambl Issues* 2012;**27**:1–21.

26 Khazaal Y, van Singer M, Chatton A, *et al.* Does Self-Selection Affect Samples' Representativeness in Online Surveys? An Investigation in Online Video Game Research. *J Med Internet Res* 2014;**16**. doi:10.2196/jmir.2759

27 Currie SR, Hodgins DC, Casey DM. Validity of the Problem Gambling Severity Index Interpretive Categories. *J Gambl Stud* 2013;**29**:311–27. doi:10.1007/s10899-012-9300-6

28 Costes J-M, Kairouz S, Eroukmanoff V, *et al.* Gambling Patterns and Problems of Gamblers on Licensed and Unlicensed Sites in France. *J Gambl Stud* 2016;**32**:79–91. doi:10.1007/s10899-015-9541-2

29 Khazaal Y, Chatton A, Bouvard A, *et al.* Internet Poker Websites and Pathological Gambling Prevention Policy. *J Gambl Stud* 2013;**29**:51–9. doi:10.1007/s10899-011-9288-3

30 Auer M, Malischnig D, Griffiths M. Is "pop-up" messaging in online slot machine gambling effective as a responsible gambling strategy? *J Gambl Issues* 2014;**29**:1–10. doi:10.4309/jgi.2014.29.3

31 Caillon J, Grall-Bronnec M, Hardouin J-B, *et al.* Online gambling's moderators: how effective? Study protocol for a randomized controlled trial. *BMC Public Health* 2015;**15**:519.