A Bayesian network model for predicting pregnancy after in vitro fertilization

G. Corani[a], C. Magli[b], A. Giusti[a], L. Gianaroli[b], L. M. Gambardella[a]

[a]*Istituto Dalle Molle Intelligenza Artificiale (IDSIA)*
Manno, Switzerland
[b]*International Institute for Reproductive Medicine (IIRM)*
Lugano, Switzerland

Abstract

We present a Bayesian network model for predicting the outcome of in vitro fertilization (IVF). The problem is characterized by a particular missingness process; we propose a simple but effective averaging approach which improves parameter estimates compared to the traditional MAP estimation. We present results with generated data and the analysis of a real data set. Moreover, we assess by means of a simulation study the effectiveness of the model in supporting the selection of the embryos to be transferred.

Keywords: In Vitro Fertilization (IVF), Bayesian networks, EM algorithm, MAP estimation, Classification

1. Introduction

According to the World Health Organization, infertility affects more than 80 million people worldwide; in vitro fertilization (IVF) is a treatment for addressing this problem. In IVF, a semen specimen is merged with a female egg in laboratory to eventually generate an *embryo*. Whenever possible, multiple embryos are cultured for each woman. Embryos are cultured for 2-5 days, before being transferred to the woman. During the culture, the morphology of each embryo is monitored at fixed time intervals; embryos with certain morphologies have indeed high implantation potential [1, 2, 3] and are

*Corresponding author
Email address: giorgio@idsia.ch (G. Corani)

thus graded as of *top* quality. Despite the effort for designing effective scoring system for the embryos [2], predicting blastocyst development remains a challenging problem [4], although promising results have been recently obtained by analyzing time-lapse embryo images collected by automated image monitoring systems [5, 6].

Reliably predicting the IVF outcome is thus still substantially an open problem [7, 8, 9]. A pioneering approach for estimating the probability of single and multiple pregnancy after an IVF treatment is the embryo - uterine model (EU) [10], which assumes that, for pregnancy to happen, it is necessary both a *receptive* uterus and a *viable* embryo. We represent *uterine receptivity* as the binary variable U, with states $\{u, \neg u\}$ (u denoting receptivity, $\neg u$ non-receptivity); we represent *embryo viability* as the binary variable E, with states $\{e, \neg e\}$ (e denoting viability, $\neg e$ non-viability).

We denote by θ_e and θ_u respectively the probabilities of the embryo to be viable and of the uterus to be receptive, namely $\theta_e = P(E = e)$, and $\theta_u = P(U = u)$. The EU model estimates the probability of pregnancy after the transfer of a *single* embryo as $\theta_e\theta_u$, thus assuming the independence of viability and receptivity. When dealing with the transfer of *multiple* embryos, each embryo is assumed to implant independently from the others. For instance, if *two* embryos are transferred, the probability of double pregnancy is $\theta_e^2\theta_u$. The *EU assumption* is therefore that pregnancy will follow only if the uterus is receptive; if this is the case, k babies will be born where k is the number of viable embryos among the transferred ones. The main limit of the original EU model is the unrealistic assumption of θ_e and θ_u being identical for respectively all embryos and all women. Therefore, in [11] the model has been reworked (adopting a generalized linear model framework) by letting vary both θ_u and θ_e on external covariates; in particular, by letting θ_u depend on the age of the woman and θ_e on the number of cells present in the embryo at a given day (this is a marker of implantation capability). More recently it has been investigated [12] how to select the number and the types of covariates on which θ_u and θ_e should depend. In fact, quantifying how uterine receptivity and embryo viability vary as a function of respectively e.g. the age of the woman or the embryo score can provide important insights to domain experts.

However, analyzing the IVF data under the EU assumption implies a *partial observability* problem. For instance, if pregnancy does *not* occur, it cannot be ascertained whether a) the uterus was *non-receptive*, b) *all* the transferred embryos were *non-viable* or c) both. If pregnancy occurs, the

uterus is known to be receptive, but it is still unknown which of the embryos gave rise to the pregnancy, unless the number of babies equals the number of transferred embryos. The missingness process is MAR (missing at random) and thus the parameters can be learned via the Expectation-Maximization (EM) algorithm [11, 8].

In the Bayesian setting, EM is typically used to identify the parameter values which maximize (although only in a local fashion) the posterior probability of the data; this is the so-called MAP (most probable a posteriori) estimation. When dealing with incomplete samples, the fully Bayesian estimation of the parameters (which requires integrating over the posterior distribution of the parameters rather than finding its maximum) is not feasible. MAP estimation is feasible also with incomplete samples, but it "*does not offer the same benefits as a full Bayesian estimation. It does not attempt to represent the shape of the posterior and thus does not differentiate between a flat posterior and a sharply peaked one. As such, it does not give us a sense of our confidence in different aspects of the parameters, and the predictions do not average out our uncertainty.*" [13, Section 17.4.4].

In a previous publication [14] we have introduced a novel probabilistic model of IVF transfers, which is a Bayesian network model based on the EU assumption. In [15] we have proposed a simple but effective averaging approach for estimating the parameters of the model from incomplete samples, which improves over the traditional MAP estimation.

In this paper we extend the analysis of [15], dealing with models which contains more variables than previously considered. Novel experiments confirm that the averaging methodology yields better parameter estimates than MAP estimation. Moreover, we compare the proposed model with state-of-the-art classification algorithms in the analysis of a data set containing IVF cycles performed at IIRM (International Institute for Reproductive Medicine) of Lugano. Eventually, we investigate via simulation the effectiveness of the model in supporting the decision of which embryos to transfer to the woman. Such a decision is typically difficult: it entails a trade-off between *maximizing* the probability of *single* pregnancy and *minimizing* the probability of *multiple* pregnancy (which is dangerous for the health of both mother and babies). In particular, we compare via simulation the outcome of the decisions taken on the basis of the model predictions and the outcome of the single-embryo transfer [16, 17].

2. The Bayesian Network model

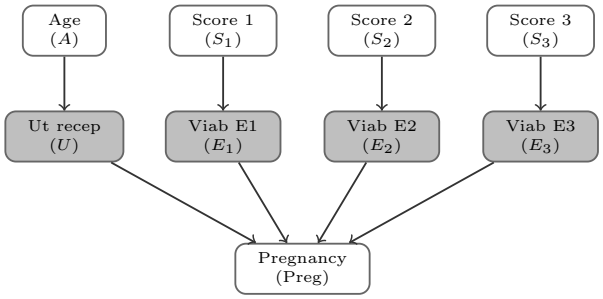


Figure 1: The BN_1 structure: nodes affected by the missingness process are shown with a gray background.

Given a generic variable X, we denote by θ_X the probability mass function which associates a marginal probability to each different value of X; we denote by $\theta_{X|Pa(X)}$ the probability mass function which associates a conditional probability to each different value of X, given each possible configuration of the parents of X, denoted as $\mathrm{Pa}(X)$. We moreover denote by $\boldsymbol{\theta}$ the set of all the parameters of the BN model.

As a first proposal, we represent the IVF transfer by the BN_1 *structure* shown in Fig. 1; a structure is a directed graph which connects the nodes representing the variables. The model manages IVF cycles with up to three embryos, as this is the maximum allowed under the Swiss law; however, it can be straightforwardly extended to manage a higher number of transferred embryos. The woman age is discretized as {<34, 34-40, 40+}.

We denote by $\mathcal{S}$ the set of nodes $\{S_1, S_2, S_3\}$; in the following, they are referred to as the $\mathcal{S}$-nodes. Such nodes take values in {*no-transfer, ntop, top, toph*} and thus represent the score of the embryos; ntop stands for non-top and toph for top-history. See Section 5.1 for more details on the meaning of the scores.

The no-transfer state allows to model cycles with less than 3 transferred embryos: in most cycles only 1 or 2 embryos are transferred in order to reduce the danger of multiple pregnancy. Notice that the different positions (1,2,3) are randomly assigned to the embryos.

The $\mathcal{S}$-nodes are *tied*: they share the same mass function θ_S instead of having separate mass functions θ_{S_1}, θ_{S_2} and θ_{S_3}. This prevents the same

embryo score (e.g., top) having a different marginal probability depending on whether one refers to node S_1, S_2 or S_3.

Node U represents uterine receptivity; it is therefore binary, with states $(u, \neg u)$.

We denote by $\mathcal{E}$ the set of nodes $\{E_1, E_3, E_3\}$, which are referred to in the following as $\mathcal{E}$-nodes. Each $\mathcal{E}$-node represents the viability of a different embryo; each $\mathcal{E}$-node is thus binary with states $(e, \neg e)$. The $\mathcal{E}$-nodes share the parameter set of the conditional mass function $\theta_{E|S}$, rather than having independent mass functions $\theta_{E|S_1}$, $\theta_{E|S_2}$ and $\theta_{E|S_3}$. Again, this prevents two embryos with the same score being given different probability of being viable just because they occupy a different position.

The pregnancy node *Preg* has four states $\{0, 1, 2, 3\}$, corresponding to the number of babies which might be born after having transferred up to three embryos. The CPT (conditional probability table) of *Preg* encodes the EU assumption; namely if the uterus is not receptive, no pregnancy will follow; if instead the uterus is receptive, k babies will be born where k is the number of viable embryos among the transferred ones. For instance, given a receptive uterus and two viable embryos out of three transferred, the CPT of node *Preg* assigns probability 1 to the outcome $P = 2$ and probability 0 to all the remaining outcomes.

2.1. The missingness process

In the following we describe the missingness process which affects receptivity and viabilities. The missingness process (MP) turns the complete data into incomplete according to a certain probability. The missingness process is MAR (*missing at random*) if the probability of a certain value to be turned into missing is independent of the value itself, although it can depend on other observed variables [18, Chap.21]. As an example, consider a clinical practice in which test A is always observed while test B is performed only if test A is positive. Thus, B is missing whenever A is negative. Given the observed outcome of A, the probability of B to be missing does not depend on the value of B itself. The missingness process is instead MCAR (*missing completely at random*) if the distribution of the missingness process is independent of both the missing and the observed values. Thus, MCAR is a particular case of MAR.

Training stage Let us consider an IVF cycle in which all the 3 embryos are transferred. At the training stage, the class variable *Preg* is always observed. In case of no-pregnancy (*Preg*=0), it is unknown whether the uterus

was not receptive, all the embryos non-viable, or both; thus the observation of both U and the $\mathcal{E}$-nodes is missing. In case of single or double pregnancy (*Preg*=1 or *Preg*=2), it can be inferred that the uterus was receptive, but it is unknown which of the embryos implanted. Thus, the $\mathcal{E}$-nodes are missing. If instead three babies are born (*Preg*=3), it can be inferred that the uterus was receptive and that all the transferred embryos were viable: both U and the $\mathcal{E}$-nodes are observed. Since the probability of E and U being missing only depends on the value of the observed variable *Preg*, the missingness process is MAR.

Test stage At test stage, we assess the ability of the model in making predictions; the *Preg* node is *always missing*. As a consequence, receptivity and viabilities are *always missing*. The MP which affects all such variables is MCAR, since the probability of being missing is identical (actually it is 1) regardless the value of observed and unobserved values. Thus, the missingness process which affects receptivity and viability is MAR at the training stage and MCAR at the test stage.

Less than three embryos transferred In most cycles less than three embryos are transferred, to reduce the danger of multiple pregnancy. For these cycles, the missingness process affects only the nodes representing the viability of the *transferred* embryos, which we denote as the $\mathcal{E}_t$-nodes. Analogously to the previous case, the $\mathcal{E}_t$-nodes are affected by a MAR missingness process in training and by a MCAR missingness process in test. The viability of non-transferred embryos is always known: a non-transferred embryo is always non-viable.

2.2. Two-parents model

The BN_2 structure, shown in Figure 2, assigns a second parent to U and to each $\mathcal{E}$-node. For U the second parent is C, namely the number of IVF *cycles* already undertaken by the woman; this variable is discretized as {0-1, >1}. A higher number of IVF cycles already undertaken by the woman is indeed a negative prognostic factor.

The second parent of each $\mathcal{E}$-node is I, namely whether or not ICSI (intracytoplasmic sperm injection) has been used. ICSI is adopted in case of severe problems with the quality of the male semen; thus, ICSI is generally correlated with a lower implantation capability of the embryo. We model ICSI as a binary variable, with states $\{i, \neg i\}$. Given the amount of missing data affecting both U and the $\mathcal{E}$-nodes, it seems unwise to further increase the number of parents of such nodes, unless a very large data set is available.

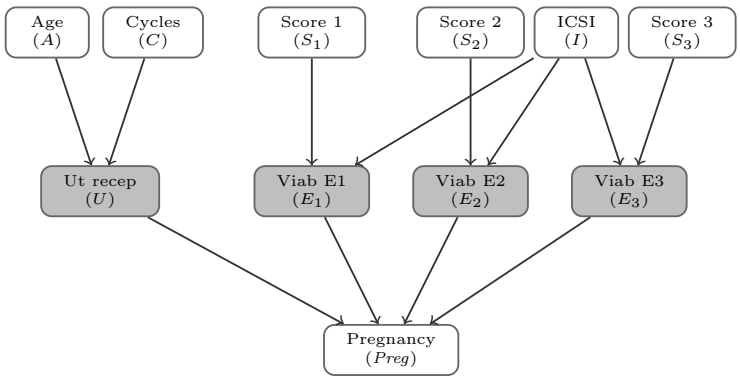


Figure 2: The BN_2 structure: nodes affected by the missingness process are shown with a gray background.

However, variables different from those proposed here can be considered as parents for U and the $\mathcal{E}$-nodes; to select the most suitable among them, in Section 5.2 we adopt a model selection procedure.

3. Estimation procedure

Given a generic variable X, we denote by θ_X^x the probability $P(X = x)$ and by $\theta_{X|Y}^{x|y}$ the probability $P(X = x|Y = y)$; this additional notation allows accessing singletons of the mass functions. We denote as $\mathcal{X}$ the set of all variables which constitute the BN model and by $\boldsymbol{x}$ an *instance*, namely the set of values and missing observations referring to the same IVF cycle. In the following for simplicity we refer to model BN_1. Let us consider the following instance $\boldsymbol{x}$ of the training set:

A	U	S_1	S_2	S_3	E_1	E_2	E_3	$Preg$
40+	*u*	*top*	*ntop*	*toph*	*?*	*?*	*?*	*1*

in which a single pregnancy has occurred; thus, the uterus is known to be receptive (U=*u*) but the observations of the $\mathcal{E}$-nodes are missing, since it is unknown which of the three embryos has implanted. Because of the incomplete samples, the likelihood needs to sum up over all the possible data completions. The possible data completions are those in which exactly one

out of three embryos is viable; the likelihood of the instance is thus:

$$
\begin{aligned}
P(\boldsymbol{x}|\boldsymbol{\theta}) = \theta_A^{40+} \cdot \theta_U^{u} \cdot \theta_S^{top} \cdot \theta_S^{ntop} \cdot \theta_S^{toph} \cdot \\
\cdot[\theta_{P|U,\mathcal{E}}^{1|u,e_1,\neg e_2,\neg e_3} \cdot \theta_{E|S}^{e|top} \cdot \theta_{E|S}^{\neg e|ntop} \cdot \theta_{E|S}^{\neg e|toph} \\
+\theta_{P|U,\mathcal{E}}^{1|u,\neg e_1,e_2,\neg e_3} \cdot \theta_{E|S}^{\neg e|top} \cdot \theta_{E|S}^{e|ntop} \cdot \theta_{E|S}^{\neg e|toph} \\
+\theta_{P|U,\mathcal{E}}^{1|u,\neg e_1,\neg e_2,e_3} \cdot \theta_{E|S}^{\neg e|top} \cdot \theta_{E|S}^{\neg e|ntop} \cdot \theta_{E|S}^{e|toph}]
\end{aligned}
$$

Notice that the likelihood contains terms θ_S^{top}, θ_S^{ntop} and θ_S^{toph} rather than $\theta_{S_1}^{top}$, $\theta_{S_2}^{ntop}$ and $\theta_{S_3}^{toph}$, since the $\mathcal{S}$-nodes share the same mass function. The same consideration applies to the $\mathcal{E}$-nodes; in the likelihood it appears e.g. $\theta_{E|S}^{\neg e|top}$ rather than $\theta_{E|S_1}^{\neg e|top}$.

The log-likelihood for the whole training set is obtained by summing the logs of the likelihood of all instances; it has a complex expression, which would be very difficult to analytically optimize. However, recalling that the missingness process at the training stage is MAR, the Expectation-Maximization algorithm (EM) can be used to maximize the likelihood. Actually, we use EM to maximize the posterior probability of the parameters (*MAP score*) rather than the likelihood, as this approach is known to reduce the danger of overfitting [19]. The MAP score can be interpreted as a kind of *penalized* likelihood.

EM can only identify a *local* maximum of the MAP score; therefore, it is common initializing the EM from m different starting points, to eventually select the estimate yielding the highest MAP score. In the following, we refer this procedure as *MAP estimation*. By definition, MAP estimation selects the most probable estimate of the parameters a posteriori. In contrast, a full Bayesian estimation would require to integrate over the posterior distribution of the parameters; such an approach cannot however be applied when dealing with incomplete samples. As already discussed, MAP estimation does not offer the same benefits as a full Bayesian estimation, as it does not attempt to represent the shape of the posterior. Thus, MAP estimation is a good approximation of Bayesian estimation when the posterior is sharply peaked around the maximum; this is however not the case when learning from incomplete samples. Typically, different EM runs achieve close values of the MAP score, returning however very different parameter estimates. Therefore, the posterior presents many local maxima rather than being sharply

peaked; MAP estimation in this context is hardly robust. This consideration remain valid even if informative starting point are adopted for EM, which allows improving the estimates.

As an alternative to MAP estimation, we propose the following averaging approach. With reference to a generic parameter θ_X^x, we average its estimates obtained in the m EM runs as follows:

$$\hat{\theta}_X^x = \frac{\sum_{i=1}^{m} \hat{\theta}_X^{x-i} P(\hat{\boldsymbol{\theta}}^{\boldsymbol{i}}|D)}{\sum_{k=1}^{m} P(\hat{\boldsymbol{\theta}}^{i}|D)} \tag{1}$$

where $\hat{\theta}_X^{x-i}$ and $P(\boldsymbol{\theta^i}|D)$ denote respectively the estimate of θ_X^x and the MAP score obtained in the i-th EM run, once it has converged. We average all parameters of the model according to Equation (1).

To illustrate the rationale of our approach, consider the general query $P(\mathcal{Z}|\boldsymbol{y}, D)$, where $\mathcal{Z}$ is the set of variables being queried, and $\boldsymbol{y}$ is the evidence available on the subset of variables $\mathcal{Y} \in \mathcal{X}$. A fully Bayesian inference would be:

$$P(\mathcal{Z}|\boldsymbol{y}, D) = \int P(\mathcal{Z}|\boldsymbol{y}, D, \boldsymbol{\theta}) P(\boldsymbol{\theta}|D) d\boldsymbol{\theta} \tag{2}$$

while, under MAP estimation, the above integral is roughly approximated as:

$$P(\mathcal{Z}|\boldsymbol{y}, D) \approx P(\mathcal{Z}|\boldsymbol{y}, D, \hat{\boldsymbol{\theta}}) \tag{3}$$

where $\hat{\boldsymbol{\theta}}$ represents the most probable parameters estimate a posteriori.

The following *pseudo-Bayesian* approach moves towards the Bayesian inference, by sampling the posterior in correspondence of the local maxima identified by the m EM runs:

$$P(\mathcal{Z}|\boldsymbol{y}, D) \simeq \sum_{i=1}^{m} P(\mathcal{Z}|\boldsymbol{y}, D, \hat{\boldsymbol{\theta}}^i) P(\hat{\boldsymbol{\theta}}^i|D) \tag{4}$$

The pseudo-Bayesian approach should generate more accurate inferences than the MAP approach, because it partially reconstructs the shape of the posterior. Yet, it requires keeping a collection of e.g. m=20 networks, each characterized by the same structure but different parameters, thus compromising the possibility for IVF experts to readily interpret the model.

The averaging idea of Equation (1) aims at keeping as much as possible the benefits of the pseudo-Bayesian approach, but instantiating only a single network. In particular, averaging the parameters according to Eq. (1) produces the same inferences of the pseudo-Bayesian approach of Eq. (4), in case of a query of type $P(X = x|\mathrm{pa}(X))$, where $pa(X)$ denotes an instantiation of all the parents of X. Only for this kind of query, the returned inference correspond to the parameter $\theta_{X|Pa(X)}^{x|pa(X)}$ of the network; averaging the parameters according to Eq. (1) is equivalent to averaging the inferences according to Eq. (4). However, a single network with parameters averaged according to Eq. (1) does not yield the same inferences than the pseudo-Bayesian approach of Eq. (4) in more general queries. In general, it is not possible replicating by a single network the inference produced by a set of networks.

Moreover, the property of parameter decomposability, which allows estimating independently the different conditional probability mass function, does not hold if the training set is incomplete [13, Chap. 19.1.3]. This prevents in principle averaging the parameters referring to the same conditional mass functions across the different EM runs. Nevertheless, the experiments of the next section show that the averaging approach consistently outperforms MAP estimation, both in parameter estimation and predictive inference, suggesting that the benefit of going towards Bayesian estimation outweighs the shortcomings of the introduced approximations.

4. Experiments with synthetic data

By dealing with generated data, it is possible analyzing the distance between the true model which has generated the data and the models which have been estimated from the incomplete samples. Adopting the notation of Section 3, given a discrete variable X, an actual probability distributions θ_X and its estimate $\hat{\theta}_X$, the distance between θ_X and $\hat{\theta}_X$ is usually measured through the Kullback-Leibler divergence (KL- divergence):

$$\mathrm{KL}(\theta_X, \hat{\theta}_X) = \sum_{x \in \Omega_X} \theta_X^x \log \frac{\theta_X^x}{\hat{\theta}_X^x}$$

where Ω_X denotes the domain of X. In case of two Bayesian networks, the KL-divergence factorizes according to the graphical structure of the network, as discussed for instance in [20].

In this section we thus report the KL-divergences and the predictions obtained using the MAP and the averaging approach. We perform 100 experiments *for each* structure (BN_1 and BN_2) and *for each* sample size $n \in \{50, 150, 300, 450, 600\}$. Each *experiment* is constituted by the following steps: a) random drawing of the parameters of the structure, thus instantiating the *true network* of the experiment; b) sampling of n complete instances from the true network; c) application of the MAR missingness process of the training stage, described in Section 2, to generate the incomplete training set (on average, U and the $\mathcal{E}$-nodes are respectively missing in 75% and 80% of the training instances), from which to estimate the parameters; d) execution of EM from m=20 different initializations (EM is stopped when the change in the value of log-likelihood among two successive iterations is smaller than 0.1%) and estimation of the parameters adopting the MAP and the averaging approach; e) evaluation of the KL-divergence between the estimated models and the true one; f) generation of the test set, by sampling further 1000 instances from the true network and removing the observations of U and the $\mathcal{E}$-nodes, thus applying the MCAR missingness process which characterizes the test stage; g) classification of the test instances. We assume to know the structure which has generated the data; namely, we only focus on the problem of parameter learning.

For this problem, AUC is a more meaningful measure than accuracy: typically, some 70-80% of the cycles ends with non-pregnancy and thus a *trivial predictor* which always returns no-pregnancy would achieve an *apparently* high accuracy of 70-80%, without however providing any useful information. The AUC overcomes this problem [21]: the AUC of the trivial predictor is indeed 0.5, while the maximum attainable AUC is 1. Since the problem has 4 classes, we measure the classification performance by computing 4 AUCs: one for each of no-pregnancy, single, double and triple pregnancy; they are denoted as $\mathrm{AUC}_0, \mathrm{AUC}_1, \mathrm{AUC}_2, \mathrm{AUC}_3$.

4.1. Results with BN_1

As shown in Figure 3, the averaging approach reduces the KL-divergence from the true network, compared to MAP estimation; the reduction of the KL-divergence is *significant* (t-test, $p < 0.01$) at *each* sample size. For n=50, the averaging approach reduces of about 52% and 62% respectively the mean and the standard deviation of the KL-divergence, compared to MAP estimation.

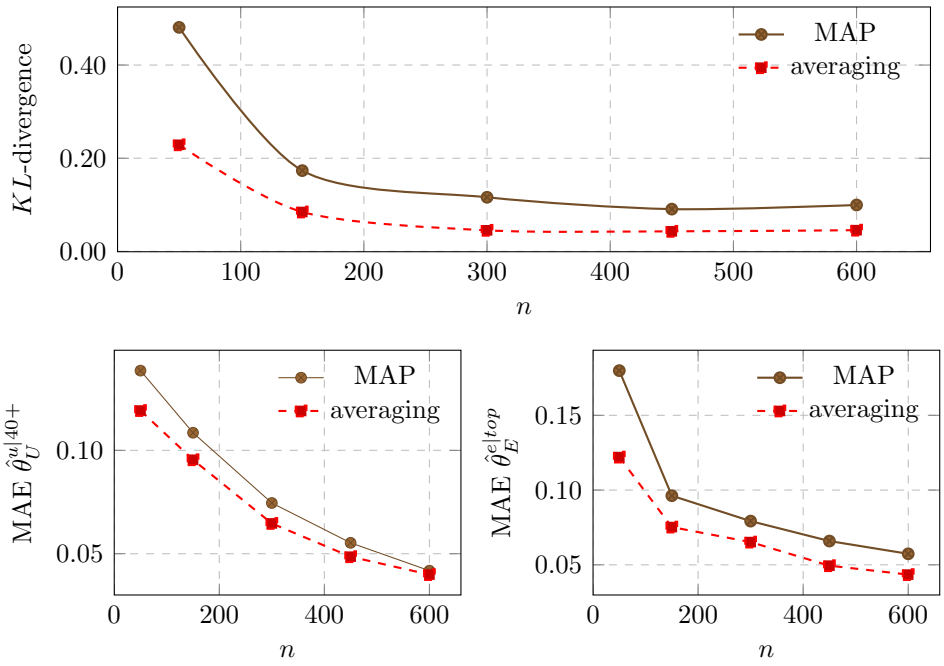


Figure 3: Experimental results on data generated with structure BN_1; the averaging approach *reduces* the KL-divergence from the true model and thus the mean absolute error (MAE) in the estimation of the parameters. Each point represents the average over 100 experiments.

The KL-divergences decrease with the sample size n, since more data allows better estimates. However, even for n=600 the averaging approach reduces the mean and the standard deviation of the KL-divergence of respectively 54% and 64%.

A lower KL-divergence from the true model implies a better estimate of the model parameters; denoting by k the number of performed experiments (k=100 in our setup), the mean absolute error (MAE) in the estimation of a generic parameter θ_X^x is:

$$\text{MAE}(\hat{\theta}_X^x) = \frac{1}{k}\sum_{i=1}^{k} |\theta_X^{i-x} - \hat{\theta}_X^{i-x}| \tag{5}$$

where θ_X^{i-x} denotes the value of θ_X^x instantiated in the true model in the i-th experiment, and $\hat{\theta}_X^{i-x}$ its estimate in the i-th experiment. It is of particular

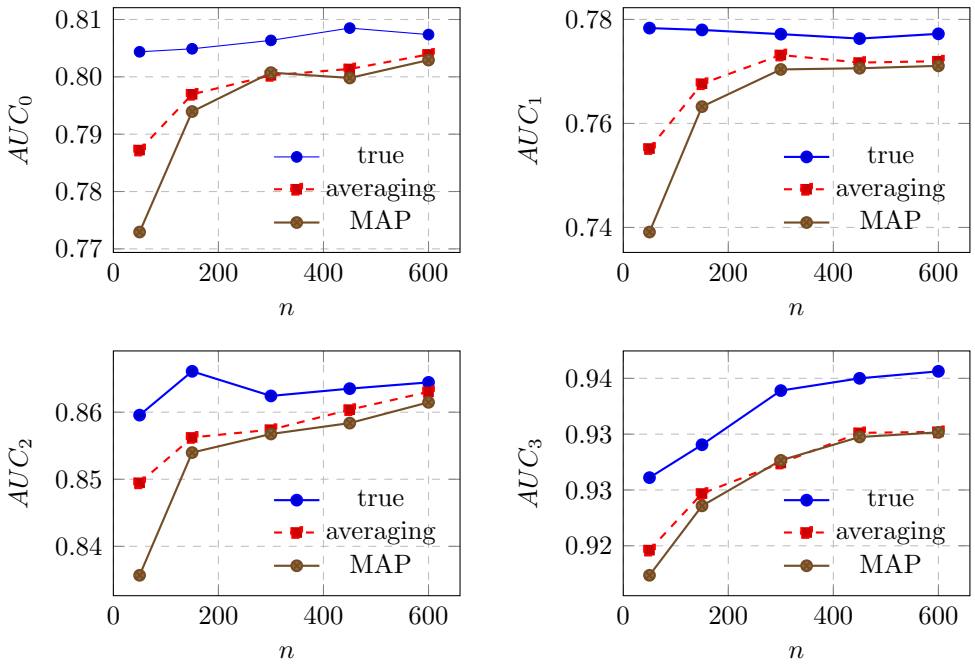


Figure 4: Experimental results on data generated with structure BN_1; the averaging approach *increases* the AUCs compared to MAP estimation. Each point represents the average over 100 experiments.

interest assessing the MAE in the estimate of parameters of the mass functions $\theta_{E|S}$ and $\theta_{U|A}$, which are severely affected by the missingness process. The lower plots of Figure 3 show, with reference to two parameters randomly chosen from such mass functions, that the averaging approach yields lower MAEs of the estimates compared to the MAP approach. For both parameters, the reduction of the MAE is significant(t-test, $p < 0.01$) at each sample size.

The improved estimates result in better classifications, as can be seen from Figure 4. The AUCs of the averaging approach are consistently higher than those of the MAP approach; considering 4 different AUCs and 5 sample sizes, there are 20 possible combinations n-AUC; in 7 out of such 20 combinations the averaging approach yields a significant (t-test, $p < 0.01$) increase of AUC compared to MAP estimation; in the remaining 13 configurations, there is no significant difference between the AUCs obtained by the averaging and

the MAP approach. Nevertheless, the improvement yielded by the averag-

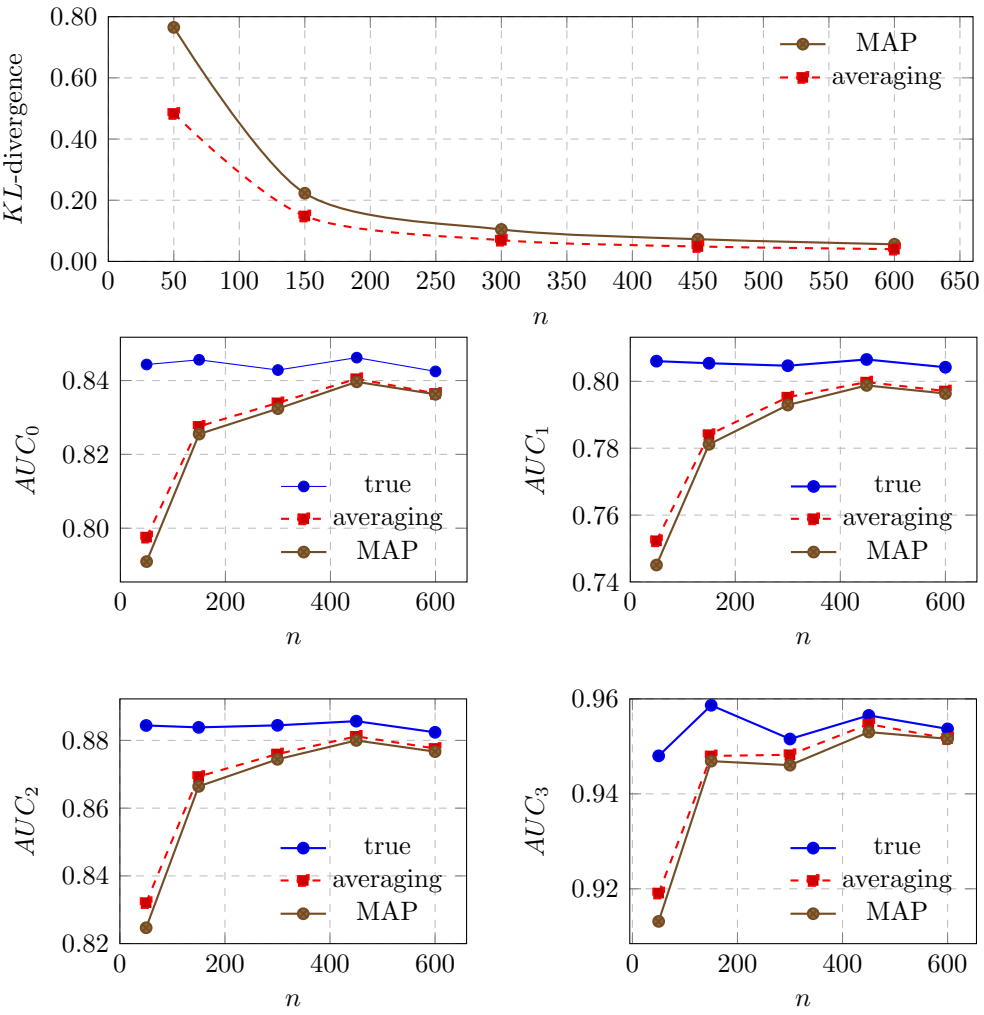


Figure 5: Experimental results on data generated with structure BN_2; the averaging approach *decreases* the KL-divergence from the true model and *increases* the AUCs compared to MAP estimation. Each point represents the average over 100 experiments.

ing approach on the AUC is less important than that on the KL-divergence. This can be partially explained by considering that the AUC cannot be increased much further, even if a perfect parameter estimate is available. In

fact, the AUC of the *true* model is generally not very far from that of the *estimated* models: the average AUC (averaging $AUC_0, AUC_1, AUC_2, AUC_3$ over all experiments) is 83.4, 83.7 and 84.5 for respectively the MAP, the averaging approach and true model. The point is that at test stage, as already discussed, U and the $\mathcal{E}$-nodes are never observed; this constitutes a major limit on predictive performance, even if the model parameters are perfectly known.

4.2. Results with BN_2

The results obtained with BN_2 are consistent with those just discussed for BN_1. The averaging approach significantly reduces (t-test, $p < 0.01$) the KL-divergence at each sample size; for instance, for n=50, it decreases the average and the standard deviation of the KL-divergence of respectively 37% and 54%; for n=600, the reduction of the mean and the standard deviation is about 28% and 32%. Note that the reductions are smaller in *percentage* compared to experiments with BN_1, but larger in *absolute* values, since the more parameterized BN_2 structure implies a larger KL-divergence of the estimated models.

Out of the 20 possible combinations n-AUC (5 values of n, 4 types of AUC) the averaging approach yields a significantly higher (t-test, $p < 0.01$) AUC than MAP estimation in 13 cases; for the remaining combinations, the difference is not significant. The AUCs of the true model are not *much* higher than those of the estimated models; as we have already discussed, a major reason of this phenomenon are the missing values of U and of the $\mathcal{E}$-nodes at test stage.

5. Analysis of the IIRM data set

We analyze 388 cycles performed at the International Institute for Reproductive Medicine (IIRM) of Lugano. Patients are unselected for age, sperm quality or any other criterion. The average age of the women is 36.3 years; the average number of transferred embryos is 2.1. ICSI is used in about 63% of the cycles and conventional IVF in the remaining 37%. The pregnancies are verified about 7 weeks after the embryo transfer. The percentage of no pregnancy, single pregnancy and double pregnancy is respectively 80%, 16% and 4%; no triple pregnancies are present in the data set.

5.1. Scoring embryos

The embryos are transferred to the woman after having been cultured for 2-5 days; the length of the culture depends on clinical and embryological considerations. During the culture, each embryo is observed and *graded* once a day; the grade is assigned by assessing the morphology of the embryo according to state of the art criteria [22, 3, 2]. Embryos whose morphology meets several criteria are graded as *top*, namely they are expected to have higher probability of implantation (if transferred into a receptive uterus) than *non-top* ones. We synthesize the grades as a binary judgment: either *top* or *non-top*. We adopt the following terminology: *grading* an embryo means assigning a top/non-top judgment; this is done on each day of the culture. Instead, *scoring* an embryo means synthesizing all the grades of a certain embryo into a single score; this is done once, at the end of the culture. We then use the scores to develop the predictive models discussed in the previous sections.

Often, the most recent grade is used as a score. However it has been suggested [23, 24, 25] that a combination of the grades obtained by the same embryo during the culture is more informative than the most recent grade. However, once an embryo has reached an *advanced* stage of culture and presents a good morphology, the chances of implantation are high irrespective of the grade at the previous stages [26]. On this basis, we synthesize the sequence of grades into a score as follows: if the most recent grade is *non-top*, the score is *non-top*; if the most recent grade is *top*, the score can be either *top* or *top-history*, depending on the previous grades. In particular, we assign the *top-history* score to a) embryos graded as top on each day of the culture and b) to embryos graded as top on each day of the culture *but one*, provided that the embryo has received at least 3 top grades (namely, the transfer is performed on day 4 or 5) and that the single non-top grade is *not* the most recent one (in which case the embryo is graded as non-top). We present some examples of embryo scoring in Table 1, for an embryo transferred on day 4.

5.2. Structure selection for the Bayesian network

We consider various structures, under the following rationale: the links between the U, the $\mathcal{E}$-nodes and *Preg* nodes is given by the EU assumption and cannot be changed. It remains instead to decide which variables to use as parents of both U and the $\mathcal{E}$-nodes. Given the partial observability problem, it seems unwise to adopt more than two parents for both U and the $\mathcal{E}$-nodes. We thus analyze multiple alternative structures which differ as for

Grades				
Day1	Day2	Day3	Day4	Score
top	top	top	top	**top-history**
top	non-top	top	top	**top-history**
top	top	non-top	top	**top-history**
top	top	top	non-top	**non-top**
top	non-top	non-top	top	**top**

Table 1: Examples of embryo scores, for a IVF transfer performed on day 4.

the parents of U and of the $\mathcal{E}$-nodes; they are listed in Table 2. Structures 1-4 refer to BN_1, BN_2 and other structures which are intermediate among the two. Structure 5 implements the idea of age impacting on embryo viability rather than on uterus receptivity, suggested in [12]. Structures 6 and 7 use the day of the transfer (discretized as {2-3, 4-5}), besides the category, to characterize the viability of the embryo: e.g., this structure allows an embryo scored as top-history on day 2-3 to have different viability from an embryo scored as top-history on day 4-5. Structures number 3 and 7 are shown in Figure 6.

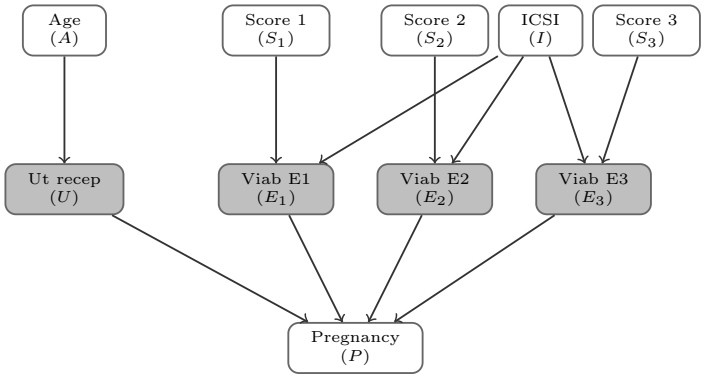
To select the most suitable structure, we use the BIC criterion [27] which, for a structure S, is defined as [13, Sec. 18.3]:

$$\mathrm{BIC} = \log P(D|S, \hat{\boldsymbol{\theta}}) - \frac{\log n}{2}|S|, \tag{6}$$

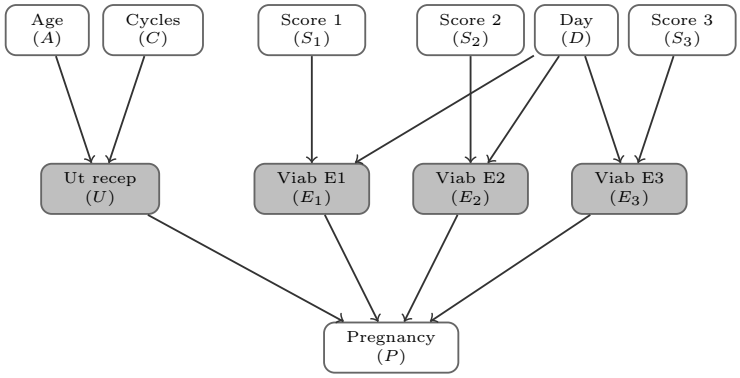
where $\log P(D|S, \hat{\boldsymbol{\theta}})$ is the log-likelihood of S evaluated in correspondence of the parameter estimate $\hat{\boldsymbol{\theta}}$ and $|S|$ is the number of free parameters of the structure. Since the data set is incomplete, the log-likelihood is maximized using the EM algorithm.

BIC trades off fit to data with model complexity; the structure which maximizes the BIC is eventually selected. We denote by ΔBIC the difference in BIC score between the highest-scoring structure and an alternative one; the value of ΔBIC can be interpreted as follows [27]:

- ΔBIC between 0 and 2: weak evidence in favor of the highest-scoring structure;
- ΔBIC between 2 and 6: positive evidence in favor of the highest-scoring structure;



(a) Candidate structure number 3



(b) Candidate structure number 7

Figure 6: Structures 3 and 7 which have been assessed on the IIRM data set.

- ΔBIC between 6 and 10: strong evidence in favor of the highest-scoring structure;
- ΔBIC >10: very strong evidence in favor of the highest-scoring structure.

We report in Table 2 the values of BIC for the various structures.

BN_1, namely the simplest structure, achieves the highest BIC score; the values of ΔBIC show strong evidence in favor of this structure compared to every considered alternative. The preference for this simple structure is reasonable if one considers the severe missingness affecting both U and the $\mathcal{E}$-nodes, which might prevent reliably estimating the parameters of more complex models.

ID	Parents of U	Parents of each $\mathcal{E}$-node	Comment	ΔBIC
1	**Age**	**Score**	**BN$_1$**	**0**
2	Age, Cycles	Score		7.5
3	Age	Score, Icsi		11.1
4	Age, Cycles	Score, Icsi	BN$_2$	18.7
5	Cycles	Score, Age		16.6
6	Age	Score, Day		23.7
7	Age, Cycles	Score, Day		31.4

Table 2: The competing structures and their ΔBIC from the highest-scoring structure, namely BN$_1$. The reported values of BIC refer to a log-likelihood computed after 20 EM initializations, using s=1 as for the equivalent sample size

The choice of structure BN$_1$ is robust. We check its sensitivity with respect to two parameters which can potentially affect the value of the likelihood and thus the BIC score: the number of EM initializations (which we let vary in {5, 10, 20} and the equivalent sample size, a parameter which controls the relative strength of prior and likelihood [13, Chap.17, pag.740], which we let vary in {1,5}. Under any choice of these parameters, BN$_1$ is the best scoring structure, with a ΔBIC of at least 6 points over the second ranked structure.

5.3. Experiments

We validate the BN$_1$ model using 10 runs of 5-folds stratified cross-validation. The BN$_1$ model achieves AUC$_0$, AUC$_1$ and AUC$_2$ which are respectively 0.741$\pm$0.06, 67.0$\pm$0.07 and 83.6$\pm$0.09. The AUCs do not significantly change if the averaging or the MAP approach are used to estimate parameters. The reported AUCs are in line with or slightly better than those reported in recent state-of-the-art studies, such as [12]. The AUCs do not significantly improve if one of the more complex structures listed in Table 2 is adopted instead of BN$_1$; this confirms the reliability of the BIC analysis.

We report in Table 3 the parameters of BN$_1$ as estimated by the averaging approach. Uterine receptivity significantly varies (Z-test, p<0.01) between each age range. Analogously, embryo viability significantly varies (Z-test, p<0.01) between each category of embryo. Our findings suggest that embryo viability is a more critical factor than uterine receptivity, as it is well

Ut. receptivity		Embryo viability	
$\theta_U^{u\|<34}$	0.78±0.04	$\theta_E^{e\|ntop}$	0.07±0.01
$\theta_U^{u\|34-40}$	0.58±0.03	$\theta_E^{e\|top}$	0.21±0.02
$\theta_U^{u\|40+}$	0.26±0.06	$\theta_E^{e\|toph}$	0.39±0.02

Table 3: Parameter estimates for the IIRM data set.

accepted within the IVF community [8, 12]. Moreover, they clearly support the introduction of the top-history score: embryos scored as top-history have significantly higher chance of implantation of embryos which are simply graded as top in the last observation. Our results thus confirm previous findings [23, 24, 25] suggesting that scores which take into consideration the sequence of grades obtained by embryo during the culture (rather than only the last grade) are more informative about the actual implantation capability of the embryo.

5.4. Comparison with traditional classifiers

Without considering the EU assumption, it is possible using a traditional classifier for predicting the IVF outcome; this is the approach followed for instance in [7] and [9]. In this section we thus compare BN_1 against traditional classifiers such as TAN, (tree-augmented naive Bayes [28]), AODE [29] (which can be seen as an ensemble of TANs), the J4.8 decision tree and the Random Forest (which is an ensemble of decision trees); see [30] and [31] for a textbook presentation of these classifier.

We build a data set containing the following features: the age of the woman; the number of IVF cycles already undertaken; the IVF method (ICSI or conventional IVF); the number of non-top, top, and top-history embryos transferred; the total number of transferred embryos. The class is the pregnancy variable with possible values {0,1,2,3}.

The AUCs are shown in Table 4. Besides AUC_0, AUC_1, AUC_2, we also report the value of the multi-class AUC, which we denote as $\overline{AUC}$. The $\overline{AUC}$ is a weighted sum of the previously mentioned AUCs, the weights being constituted by the prior probability of the various classes (no pregnancy: 80%; single pregnancy: 16%; double pregnancy: 4%).

BN_1 and AODE clearly outperform in decreasing order TAN, Random Forest and J4.8. For the case of AUC_0, we also report in Figure 7 the

ROC curve, which therefore assesses the model specificity and sensitivity in discriminating between non-pregnancy and pregnancy (either single or multiple).

	BN_1	AODE	TAN	Rand F.	J4.8
AUC_0	74.1	**74.8**	73.0	66.3	58.6
AUC_1	67.0	**68.0**	65.1	56.1	52.5
AUC_2	**83.4**	81.6	67.7	68.4	68.4
$\overline{AUC}$	73.2	74.2	71.5	64.5	58.2

Table 4: Comparison of different classifiers over the IIRM data set; the AUCs are measured by 10 runs of 5-folds cross-validation.

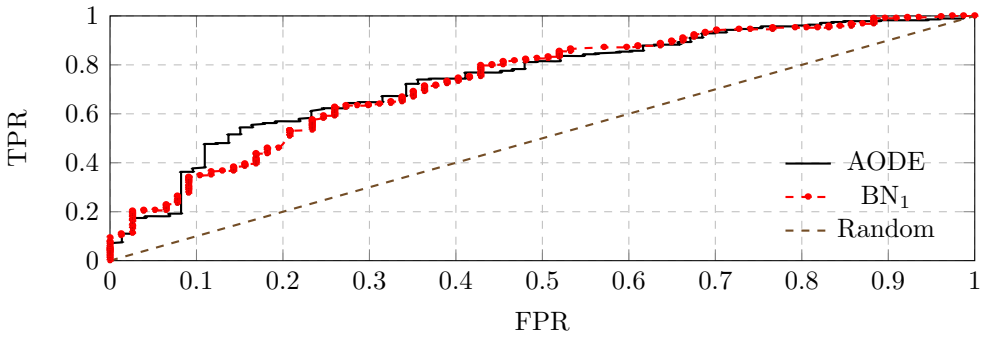


Figure 7: ROC curves obtained for the AODE and the BN_1 model, using the no-pregnancy class as the positive one; the area underlying this curve corresponds therefore to AUC_0. The bisector represents the ROC of a random guesser. TPR stands for true positive rate and FPR for false positive rate.

Thus, BN_1 has comparable performance to the most effective machine learning algorithms; unlike them, it is however interpretable from a clinical viewpoint. For this reason, we conclude that it should be preferred over them, as recognized also in [8]: "*Most clinical studies of embryo-level factors avoid the partial observability problem [...]. Modelling approaches which correctly incorporate the structure of the data are however to be preferred for the analysis of studies of embryo-level viability predictors.*"

6. Supporting decisions: a simulation study

After the culture of the embryos, it is necessary to select the embryos to be transferred to the woman, addressing the trade-off between *maximizing* the probability of *single* pregnancy and *minimizing* the probability of *multiple* pregnancy. The risks associated with a multiple pregnancy comprise for instance cognitive and physical impairment of the baby. Transferring a single embryo (single-embryo transfer, SET) prevents multiple pregnancy, but it also increases the probability of no-pregnancy [16, 17]. To have higher chance of pregnancy, many couples are available to accept some risk of multiple pregnancy [32, 17]. In this situation, prognostic models could prove very useful, providing predictions on the basis of which couples and clinicians could decide how many and which embryos to transfer; we call this policy decision-supported transfer (DST). In the following, we compare SET and DST by a simulation study.

To simulate DST we adopt maximization of the expected utility as a criterion for taking decisions [33]. Let us denote as u_i the utility of i babies being born; thus, u_0, u_1, u_2, u_3 denote respectively the utility of no-pregnancy, single, double and triple pregnancy. We define two utility functions, a *double-averse* and a *double-tolerant* one:

	u_0	u_1	u_2	u_3
double-tolerant	0	1	0.5	-1
double-averse	0	1	-1	-2

Both utility functions assign utility 0 to no-pregnancy (baseline outcome) and 1 to single pregnancy, which is the most desirable outcome. The two functions differ in the evaluation of the double pregnancy, which is regarded as partially positive by the double-tolerant utility but as strongly negative by the double-averse one. Both functions (but the double-averse in particular) are severely averse to triple pregnancy.

Each embryo can be transferred or not; three embryos (the ones that according to the national regulations could be left in culture) thus generate $2^3 = 8$ different *transfer options*, obtained by combining in all possible ways the decision of transferring or not each embryo. For each transfer option, the predictive model computes the probability of no-pregnancy, single, double and triple pregnancy. From a statistical viewpoint, the problem can be casted as one of taking optimal decision in a cost-sensitive environment [34].

However, since it is not possible assigning a monetary value to a pregnancy, the outcome of the decisions is assessed in terms of utility rather than costs.

Let us assess for instance a transfer option with the following probability of no-pregnancy, single, double and triple pregnancy: $\{.6, .2, .1, .1\}$. The expected utility of such a transfer option, under the double-tolerant utility, is $0 \cdot .6 + 1 \cdot .2 + .5 \cdot .1 - 1 \cdot .1 = -0.05$. The expected utility is computed also for all the remaining transfer options; the transfer option with highest expected utility is eventually identified as *optimal*. The identification of the optimal transfer option is independently carried out under the double-tolerant and the double-averse utility.

In the following, we use the term *probability of pregnancy* as a shorthand for the probability distribution over the outcomes of no-pregnancy, single, double and triple pregnancy. Note that uterus receptivity and embryo viabilities are *unknown* when evaluating the probability of pregnancy of each transfer option. As shown by the simulations of Section 4, this introduces a substantial uncertainty on the prediction even if the correct model is available. However to make the experiment more realistic we further introduce a model *misspecification*: we compute the probability of pregnancy using BN_1 but we simulate the outcome of the IVF cycle using BN_2, which contains some variables ignored by BN_1. In fact, there is always a mismatch between the model adopted to support decisions and the real phenomenon being modelled.

We compare SET and DST via simulation as follows:

1. sample from BN_2 the *clinical characteristics* of the cycle, namely age of the woman (A), number of cycles already undertaken (C), icsi/ conventional IVF (I);
2. sample from BN_2 the score of three embryos (S_1, S_2, S_3) and list the eights possible transfer options;
3. compute the probability of pregnancy for each transfer option using BN_1;
4. identify the optimal transfer under the double-tolerant utility, the double-averse utility and SET (note: for SET, the optimal transfer is the transfer of the best-scoring embryo).
5. simulate the outcome of the cycle using BN_2, in correspondence of the transfer options selected by the different policies.

We simulate 5000 IVF cycles; results are reported in Tables 5 and 6 and provide different interesting insights.

Policy → **Indicators ↓**	**SET**	**DST** **double-averse**	**DST** **double-tolerant**
transferred embryos	1.0(±0.00)	1.6(±0.02)	2.8(±0.02)
no-pregnancy (%)	83.1(±1.0)	81.0(±1.0)	75.1(±1.2)
single pregnancy (%)	16.9(±1.1)	18.1(±1.3)	21.0(±1.1)
double pregnancy (%)	0.0(±0.0)	0.9(±0.2)	3.7(±0.6)
triple pregnancy (%)	0.0(±0.0)	0.0(±0.0)	0.2(±0.0)

Table 5: Pregnancy rates for different policies, with 95% CIs into parentheses.

Policy → **Indicators ↓**	**SET**	**DST** **double-averse**	**DST** **double-tolerant**
non-top (%)	13.8(±1.1)	46.0(±1.4)	50.8(±1.4)
top (%)	22.2(±1.1)	16.9(±1.1)	19.2(±1.1)
top-history (%)	64.0(±1.3)	37.1(±1.3)	30.0(±1.3)

Table 6: Types of the embryos chosen for the transfer.

Let us start by comparing SET with DST under the double-tolerant utility. DST increases both single pregnancy and double-pregnancy rate of about 4 points. There is also a very small increase of the rate of triple pregnancies: from 0 under SET to 0.2% under DST. DST is however effective at controlling the rate of triple pregnancies. By transferring *all* the three available embryos in each cycle, a triple pregnancy rate of 0.6% (not reported in tables) is observed. Compared to this policy, DST reduces the average number of transferred embryos from 3 to 2.8; this reduction albeit small is very effective, resulting in a three-fold reduction of the triple pregnancy rate (from 0.6% to 0.2%).

SET implies transferring only one embryo per cycle; as a result, it can select high-quality embryos to be transferred; in particular 65% of the embryos transferred under SET are top-history. This percentage drops to only 30% for DST, as a result of both the need for transferring more embryos and the limited amount of top-history embryos. This suggest that even higher

increase of pregnancy rate could be obtained by DST, if more embryos to choose from were available; this could be the case of some countries in which it is allowed to culture up to 5 embryos.

But eventually is DST preferable to SET for a double-tolerant couple? We measure the satisfaction about the IVF outcome by the *average utility*, that is we analyze the 4000 simulated cycles assigning [0, 1, 0.5, -1] points for each [no-, single, double, triple] pregnancy; eventually, we average the score. For the double-tolerant case, DST increases the average utility of about 30% compared to SET.

The situation is quite different if DST is applied under the double-averse utility, in which case the average utility obtained by DST and SET is substantially the same. This could be expected: since the utility function is strongly averse to double pregnancy, DST cannot perform very differently from SET.

Summing up DST can importantly increase, compared to SET, the satisfaction of couples about IVF treatment if the couple is available to accept at least a moderate risk of double pregnancy. If the couple is strongly averse to double pregnancy (as in the case of the double-averse utility), DST yields however little advantage over SET.

7. Conclusions

We have introduced a Bayesian Network model of IVF based on the EU assumption. IVF cycles are characterized by partial observability of some variables; we have presented an averaging strategy which yields more reliable parameter estimates than the traditional MAP estimation. The proposed model is equally or more predictive than well-recognized classification algorithms, with the further advantage of being biologically interpretable. Analysis of the model parameters indicates that embryo viability is a more critical factor than uterus receptivity, in agreement with previous studies; it also shows the effectiveness of the top-history score for detecting embryos with high implementation potential.

In future works, the model could be evolved by scoring the embryos through time-lapse image analysis [6, 5] rather than through static observation of the embryos.

From a different viewpoint, it could be interesting refining the model selection step adopting more sophisticated procedures such as those described in [13, Chapter 19.4].

Acknowledgments

Research partially supported by CTI (Commission for Technology and Innovation) grant n. 9707.1 PFSL-LS, Swiss NSF grant no. 200020-132252 and the Hasler foundation grant n. 10030.

References

[1] Scott, L., Finn, A., O'Leary, T., McLellan, S., Hill, J.. Morphologic parameters of early cleavage-stage embryos that correlate with fetal development and delivery: prospective and applied data for increased pregnancy rates. Human Reproduction 2007;22(1):230–240.

[2] Balaban, B., Brison, D., Calderón, G., Catt, J., Conaghan, J., Cowan, L., et al. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. Human Reproduction 2011;26(6):1270–1283.

[3] Gianaroli, L., Magli, M., Ferraretti, A., Lappi, M., Borghi, E., Ermini, B.. Oocyte euploidy, pronuclear zygote morphology and embryo chromosomal complement. Human Reproduction 2007;22(1):241–249.

[4] Uyar, A., Bener, A., Ciray, H.N., Bahceci, M.. Bayesian networks for predicting IVF blastocyst development. In: 20th International Conference on Pattern Recognition (ICPR 2010). 2010, p. 2772–2775.

[5] Wong, C., Loewke, K., Bossert, N., Behr, B., De Jonge, C., Baer, T., et al. Non-invasive imaging of human embryos before embryonic genome activation predicts development to the blastocyst stage. Nature Biotechnology 2010;28(10):1115–1121.

[6] Meseguer, M., Rubio, I., Cruz, M., Basile, N., Marcos, J., Requena, A.. Embryo incubation and selection in a time-lapse monitoring system improves pregnancy outcome compared with a standard incubator: a retrospective cohort study. Fertility and Sterility 2012;98(6):1481–1489.

[7] Saith, R., Srinivasan, A., Michie, D., Sargent, I.. Relationships between the developmental potential of human in-vitro fertilization embryos and features describing the embryo, oocyte and follicle. Human Reproduction Update 1998;4(2):121–134.

[8] Roberts, S.A.. Models for assisted conception data with embryo-specific covariates. Statistics in Medicine 2007;26(1):156–170.

[9] Morales, D., Bengoetxea, E., Larrañaga, P., García, M., Franco, Y., Fresnada, M., et al. Bayesian classification for the selection of in vitro human embryos using morphological and clinical data. Computer Methods and Programs in Biomedicine 2008;90(2):104–116.

[10] Speirs, A., Lopata, A., Gronow, M., Kellow, G., Johnston, W.. Analysis of the benefits and risks of multiple embryo transfer. Fertility and Sterility 1983;39(4):468–471.

[11] Zhou, H., Weinberg, C.. Evaluating effects of exposures on embryo viability and uterine receptivity in vitro fertilization. Statistics in Medicine 1998;17(14):1601–1612.

[12] Roberts, S., Hirst, W., Brison, D., Vail, A., et al. Embryo and uterine influences on IVF outcomes: an analysis of a UK multi-centre cohort. Human Reproduction 2010;25(11):2792–2802.

[13] Koller, D., Friedman, N.. Probabilistic Graphical Models: Principles and Techniques. The MIT Press; 2009.

[14] Gianaroli, L., Magli, M., Gambardella, L., Giusti, A., Grugnetti, C., Corani, G.. Objective way to support embryo transfer: a probabilistic decision. Human Reproduction 2013;28(5):1210–1220.

[15] Corani, G., Magli, C., Giusti, A., Gianaroli, L., Gambardella, L.. A Bayesian network model for predicting the outcome of in vitro fertilization. In: Proc. 6th European Workshop on Probabilistic Graphical Models (PGM 2012). 2012, p. 75–82.

[16] Roberts, S., Fitzgerald, C., Brison, D.. Modelling the impact of single embryo transfer in a national health service ivf programme. Human Reproduction 2009;24(1):122–131.

[17] Scotland, G., McNamee, P., Peddie, V., Bhattacharya, S.. Safety versus success in elective single embryo transfer: womens preferences for outcomes of in vitro fertilisation. BJOG: An International Journal of Obstetrics & Gynaecology 2007;114(8):977–983.

[18] Gelman, A., Carlin, J., Stern, H., Rubin, D.. Bayesian data analysis. 2004.

[19] Lauritzen, S.. The EM algorithm for graphical association models with missing data. Computational Statistics & Data Analysis 1995;19(2):191–201.

[20] Tong, S., Koller, D.. Active learning for parameter estimation in bayesian networks. In: Proc. Neural Information Processing Systems (NIPS). 2000, p. 647–653.

[21] Bradley, A.. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997;30(7):1145–1159.

[22] Volpes, A., Sammartano, F., Coffaro, F., Mistretta, V., Scaglione, P., Allegra, A.. Number of good quality embryos on day 3 is predictive for both pregnancy and implantation rates in in vitro fertilization/intracytoplasmic sperm injection cycles. Fertility and Sterility 2004;82(5):1330–1336.

[23] De Placido, G., Wilding, M., Strina, I., Alviggi, E., Alviggi, C., Mollo, A., et al. High outcome predictability after IVF using a combined score for zygote and embryo morphology and growth rate. Human Reproduction 2002;17(9):2402–2409.

[24] Lan, K., Huang, F., Lin, Y., Kung, F., Hsieh, C., Huang, H., et al. The predictive value of using a combined Z-score and day 3 embryo morphology score in the assessment of embryo survival on day 5. Human Reproduction 2003;18(6):1299–1306.

[25] Brezinova, J., Oborna, I., Svobodova, M., Fingerova, H.. Evaluation of day one embryo quality and IVF outcome: a comparison of two scoring systems. Reproductive Biology and Endocrinology 2009;7(1):9.

[26] Guerif, F., Le Gouge, A., Giraudeau, B., Poindron, J., Bidault, R., Gasnier, O., et al. Limited value of morphological assessment at days 1 and 2 to predict blastocyst development potential: a prospective study based on 4042 embryos. Human Reproduction 2007;22(7):1973–1981.

[27] Raftery, A.. Bayesian model selection in social research. Sociological Methodology 1995;25:111–164.

[28] Friedman, N., Geiger, D., Goldszmidt, M.. Bayesian network classifiers. Machine Learning 1997;29(2):131–163.

[29] Webb, G., Boughton, J., Wang, Z.. Not so naive Bayes: aggregating one-dependence estimators. Machine Learning 2005;58(1):5–24.

[30] Witten, I.H., Frank, E.. Data Mining: Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann; 2005.

[31] Tan, P., Steinbach, M., Kumar, V.. Introduction to Data Mining. 2006.

[32] Hartshorne, G., Lilford, R.. Different perspectives of patients and health care professionals on the potential benefits and risks of blastocyst culture and multiple embryo transfer. Human Reproduction 2002;17(4):1023–1030.

[33] Parmigiani, G.. Modeling in Medical Decision Making: A Bayesian Approach. Wiley; 2002.

[34] Elkan, C.. The foundations of cost-sensitive learning. In: Proc. Int. Joint Conference on Artificial Intelligence (IJCAI - 01). 2001, p. 973–978.