Prediksi Tingkat Fertilitas Pria Dengan Algoritma Pohon Keputusan Cart

Ulfa Khaira*[1], Norman Syarief[1], Zalman[1], Isra Hayati[1]

[1]Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Jambi

*Ulfa.ilkom@gmail.com

ABSTRAK

Latar Belakang: Masalah infertilitas antara pasangan suami isteri menjadi masalah penting yang dapat mengganggu keharmonisan rumah tangga, banyak orang masih menganggap infertilitas atau ketidaksuburan sebagai masalah wanita. Namun, sekitar 7% pria di usia produktif mengalami infertiltitas. Faktor terbesar penyebab infertilitas bagi pria adalah masalah kualitas sperma. Analisa sperma dapat menjadi prediktor terbaik dalam melihat potensi fertilitas pria. Teknik machine learning dan data mining dapat digunakan dalam automatisasi diagnosa penyakit.

Tujuan: Penelitian ini bertujuan untuk mendapatkan model klasifikasi yang berupa aturan-aturan dari data sampel sperma 100 sukarelawan. Dari model klasifikasi tersebut dapat digunakan untuk melakukan prediksi tingkat fertilitas pria yang dikelaskan menjadi 2 yaitu kelas normal dan kelas altered (fertilitas menurun).

Metode: Penelitian ini menggunakan data sampel sperma 100 sukarelawan berusia antara 18 hingga 36 tahun yang dianalisis sesuai dengan kriteria WHO. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan praproses data yang meliputi pembersihan dan transformasi data. Proses datamining dilakukan dengan menggunakan algoritma CART, pada tahapan ini akan dicari pola atau informasi menarik dari data analisis sperma untuk menghasilkan model berupa pohon keputusan.

Hasil penelitian: Algoritma classification and regression trees (CART) mampu menghasilkan pohon keputusan dengan tingkat akurasi sebesar 84%. Terdapat 12 aturan dalam memprediksi tingkat fertilitas pria.

Simpulan: Dengan tingkat akurasi sebesar 84%, model pohon keputusan yang dihasilkan dalam penelitian ini dapat digunakan untuk membantu dokter dalam prediksi tingkat fertilitas pria.

Kata kunci: Algoritma classification and regression trees (CART), Fertilitas, Pria

ABSTRACT

Background: The problem of infertility between married couples becomes an important problem that can disrupt relationship harmony in the marriage, many people still regarding infertility as a woman's problem. However, about 7% of men of reproductive age have experienced fertility problems. The main causes of male infertility is sperm quality. Sperm analysis can be the best predictor in seeing the potential for male fertility. Machine learning and data mining techniques can be used in automating disease diagnoses.

Purpose: This study aims to obtain a classification model in the form of rules from sperm sample data of 100 volunteers. From the classification model it can be used to predict the level of male fertility which is classified into 2 namely normal and altered classes (decreased fertility).

Method: This study uses sperm sample data from 100 volunteers aged between 18 and 36 years who were analyzed according to WHO criteria. Before the data mining process can be carried out, it is necessary to pre-process the data which includes cleaning and transforming the data. The datamining process is carried out using the CART algorithm, at this stage interesting patterns or information from sperm analysis data will be sought to produce a model in the form of a decision tree.

Result: Classification and regression trees (CART) algorithm is able to produce decision trees with 84% accuracy. There are 12 rules in predicting male fertility.

Conclusion: With an accuracy rate of 84%, the decision tree model produced in this study can be used to assist doctors in predicting male fertility.

Keywords: Classification and regression trees (CART) algorithm, Fertility, Male

PENDAHULUAN

Dalam dua dekade terakhir ini masalah infertilitas antara pasangan suami isteri menjadi masalah penting yang dapat mengganggu keharmonisan rumah tangga (Gil, Girela, De Juan, Gomez-Torres, & Johnsson, 2012). Menurut badan kesehatan dunia WHO, tingkat fertilitas menurun drastis. Infertilitas didefinisikan sebagai ketidakmampuan untuk hamil setelah satu tahun sering melakukan hubungan seksual tanpa menggunakan metode kontrasepsi. Infertilitas dapat dikatakan sebagai penyakit pada sistem reproduksi (WHO, 2016).

Meski banyak orang masih menganggap infertilitas atau ketidaksuburan sebagai masalah wanita, namun menurut penelitian Krausz (2011), sekitar 7% pria di usia produktif mengalami infertiltitas. Faktor terbesar penyebab infertilitas bagi pria adalah masalah kualitas sperma. Kualitas sperma tidak saja dipengaruhi oleh ketidakseimbangan hormon atau masalah fisik tetapi juga masalah psikologis dan perilaku. Faktor lingkungan dan gaya hidup seperti mengkonsumsi alkohol dan rokok juga mempengaruhi tingkat kualitas sperma (semen) (Giwercman & Giwercman, 2011). Analisa sperma dapat menjadi prediktor terbaik dalam melihat potensi fertilitas pria.

Beberapa tahun terakhir ini, teknik *machine learning* dan *data mining* digunakan dalam automatisasi diagnosa penyakit (Foster, Koprowski, & Skufca, 2014). Beberapa penelitian terkait prediksi tingkat fertilitas pria telah dilakukan, Susanto (2013) melakukan komparasi algoritma Naive Bayes dan C4.5 untuk memprediksi tingkat fertilitas pria, algoritma Naive Bayes menghasilkan akurasi sebesar 83,75% sedikit lebih tinggi dari akurasi algoritma C4.5 sebesar 81,25%, Irawan (2017) melakukan prediksi kesuburan dengan menggunakan principal component analysis kemudian diklasifikasikan menggunakan metode Naïve Bayes Classifier menghasilkan akurasi sebesar 80%.

Penelitian ini menggunakan data yang sama dengan penelitian sebelumnya yaitu bersumber dari dataset UCI (University of California) Machine Learning Repository. Menerapkan teknik pohon keputusan algoritma *Classifcation and Regression Tree* (CART). Pohon keputusan merupakan salah satu metode dalam klasifikasi, metode ini juga dimanfaatkan untuk prediksi dengan menggunakan struktur pohon (Han, Kamber, & Pei, 2012). Pohon keputusan ini sebagai alat untuk mendukung pengambilan keputusan, dengan melakukan *break down* proses pengambilan keputusan yang kompleks sehingga menjadi lebih simpel. Pohon keputusan ini merupakan pendekatan yang *powerful* dalam *data mining* karena dapat menemukan pengetahuan yang bermanfaat dari data yang kompleks(Bhargava, Sharma, Bhargava, & Mathuria, 2013).

Penelitian ini bertujuan untuk mendapatkan model klasifikasi yang berupa aturan-aturan dari data sampel sperma 100 sukarelawan yang dianalisis sesuai dengan kriteria WHO 2010. Konsentrasi sperma berhubungan dengan data sosio-demografis, faktor lingkungan, status

kesehatan, dan kebiasaan hidup. Dari model klasifikasi tersebut dapat digunakan untuk melakukan prediksi tingkat fertilitas pria yang dikelaskan menjadi 2 yaitu kelas normal dan kelas altered (fertilitas menurun).

METODE

Dataset

Dalam penelitian ini dataset yang digunakan adalah data sampel sperma 100 sukarelawan mahasiswa Universitas Alicante Spanyol berusia antara 18 hingga 36 tahun yang dianalisis sesuai dengan kriteria WHO. Dataset didapatkan dari UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/datasets/Fertility#). Dataset tersebut terdiri dari 10 atribut yang digunakan untuk memprediksi tingkat fertilitas pria.

Praproses Data

Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning seperti membuang duplikasi data, memeriksa data yang inkonsisten, *replace missing values* dan memperbaiki kesalahan pada data. Pada data ini tidak ditemukan *missing value* jadi tidak perlu dilakukan *cleaning data*.

Tahapan selanjutnya adalah proses transformasi data yaitu dengan mengubah skala data asli menjadi bentuk lain, sehingga data tersebut sesuai untuk proses *data mining*.

Tabel 1. Atribut Kondisi Fertilitas Pria

Atribut	Kriteria	Nilai Bobot
Musim dimana analisis dilakukan	1)Musim dingin, 2)Musim semi, 3)Musim panas, 4) Musim gugur	(-1, -0.33, 0.33, 1)
Umur pada saat analisis	18-36 tahun	Skala [0-1]
Penyakit anak (yaitu cacar air, campak, gondok, polio)	Ya, tidak	(0,1)
Kecelakaan atau trauma serius	Ya, tidak	(0,1)
Mengalami pembedahan	Ya, tidak	(0,1)
Demam tinggi pada 1 tahun terakhir	kurang dari tiga bulan yang lalu , lebih dari tiga bulan yang lalu , tidak pernah	(-1 , 0, 1)
Frekuensi konsumsi alkohol	beberapa kali sehari, setiap hari, beberapa kali dalam seminggu, seminggu sekali, hampir tidak pernah, atau tidak pernah sama sekali	(0, 0.2, 0.4, 0.6, 0.8, 1)
Kebiasaan merokok	tidak pernah, sesekali, setiap hari	(-1, 0, 1)
Jumlah jam yang dihabiskan duduk per hari	antara 1 hingga 16 jam	Skala [0-1]
Diagnosis	Normal (N) , Altered (O)	1,0

Model Klasifikasi

Pada tahapan ini akan dicari pola atau informasi menarik dari data analisis sperma menggunakan algoritma CART untuk menghasilkan model berupa pohon keputusan, dengan skema pembentukan 75% dari data digunakan sebagai training set dan 25% digunakan sebagai testing set. Pohon keputusan yang dihasilkan CART merupakan pohon biner dimana tiap simpul wajib memiliki dua cabang. Adapun langkah-langkah untuk menghasilkan pohon keputusan dengan algoritma CART, sebagai berikut (Rutkowski, Jaworski, Pietruczuk, & Duda, 2014):

- Menyusun calon cabang seluruh variable predictor secara lengkap.
- Menilai kinerja keseluruhan calon cabang dengan cara menghitung nilai besaran kesesuaian *(goodness)*, calon cabang yang memiliki nilai *goodness* paling besar akan menjadi cabang.
- Menggambar percabangan, jika tidak ada lagi node keputusan maka pelaksanaan algoritma CART dihentikan, ulangi langkah kedua jika masih terdapat node keputusan,dengan terlebih dahulu membuang calon cabang yang telah menjadi cabang.
- Pemangkasan (Prunning) dilakukan untuk mengurangi pohon klasifikasi yang sangat besar agar menjadi sedehana.
- membentuk pohon keputusan dan melakukan pembentukan rule.

Evaluasi Model Klasifikasi

Untuk merepresentasikan akurasi dapat menggunakan Confusion matrix, menurut Kuncheva (2014) confusion matrix digunakan untuk menunjukkan persebaran galat pada seluruh kelas yang digunakan pada sebuah proses klasifikasi. Contoh confusion matrix dapat disajikan pada Tabel 2. Akurasi dari klasifikasi dapat diukur dengan menghitung jumlah data yang terklasifikasi dengan benar dibagi dengan jumlah data seluruhnya. Confusion matrix juga memperlihatkan informasi lainnya seperti kelas yang sering salah diklasifikasi. Pada Tabel 2, terlihat bahwa banyak data kelas A yang salah diklasifikasi dengan kelas B.

Tabel 2. Contoh *confusion matrix* yang menampilkan akurasi proses klasifikasi yang melibatkan 20 data dari tiga kelas (A dan B)

Kelas Aktual	Kelas Prediksi	
	A	B
A	4	7
B	0	9

Lingkungan Pengembangan

Perangkat lunak yang akan digunakan dalam penelitian ini adalah:
- Scikit-learn untuk Python 3.5
- Jupyter Notebook
- Microsoft Office Excel

PEMBAHASAN

Data yang digunakan sebagai *input* adalah data sampel sperma 100 sukarelawan berusia antara 18 hingga 36 tahun yang dianalisis sesuai dengan kriteria WHO terdiri dari 9 atribut input yang berjumlah 100 *record* yang akan dibagi ke dalam 2 bagian, yaitu 75% untuk data *training* yang akan menghasilkan suatu rule, 25% sisanya digunakan untuk data *testing*. Data ini untuk menguji hasil rule yang telah diperoleh dari proses *training*. Data hasil

praproses disajikan pada Gambar 1. Dari data tersebut, atribut ke 1 hingga 9 telah berjenis numerik, namun untuk target set yaitu atribut diagnosis, data masih berbentuk huruf (N dan O), oleh karena itu, agar proses data mining dapat dilakukan, perlu dilakukan transformasi data menjadi numerik (N = 0, dan O = 1).

	season	age	child diseases	accident or trauma	surgical intervention	high fevers in last year	alcohol consumption	smoking habit	hours spent sitting	diagnosis	target
0	-0.33	0.69	0	1	1	0	0.8	0	0.88	N	0
1	-0.33	0.94	1	0	1	0	0.8	1	0.31	O	1
2	-0.33	0.50	1	0	0	0	1.0	-1	0.50	N	0
3	-0.33	0.75	0	1	1	0	1.0	-1	0.38	N	0
4	-0.33	0.67	1	1	0	0	0.8	-1	0.50	O	1

Gambar 1. Hasil Praproses Data

Data hasil praproses siap digunakan untuk proses klasifikasi sehingga akan didapatkan model klasifikasi. Setelah model berhasil dibuat, kita dapat menguji akurasi model menggunakan data baru, yaitu data pada data testing. Akurasi direpresentasikan dengan menggunakan *confusion matrix*. Akurasi dari klasifikasi dapat diukur dengan menghitung jumlah data yang terklasifikasi dengan benar dibagi dengan jumlah data seluruhnya. Berdasarkan eksperimen yang ditampilkan melalui confusion matrix (Tabel 3), hasil akurasi dari model yang diusulkan sebesar 84%, nilai akurasi ini sedikit lebih tinggi jika dibandingkan dengan penelitian yang dilakukan susanto (2013) dan Irawan (2017).

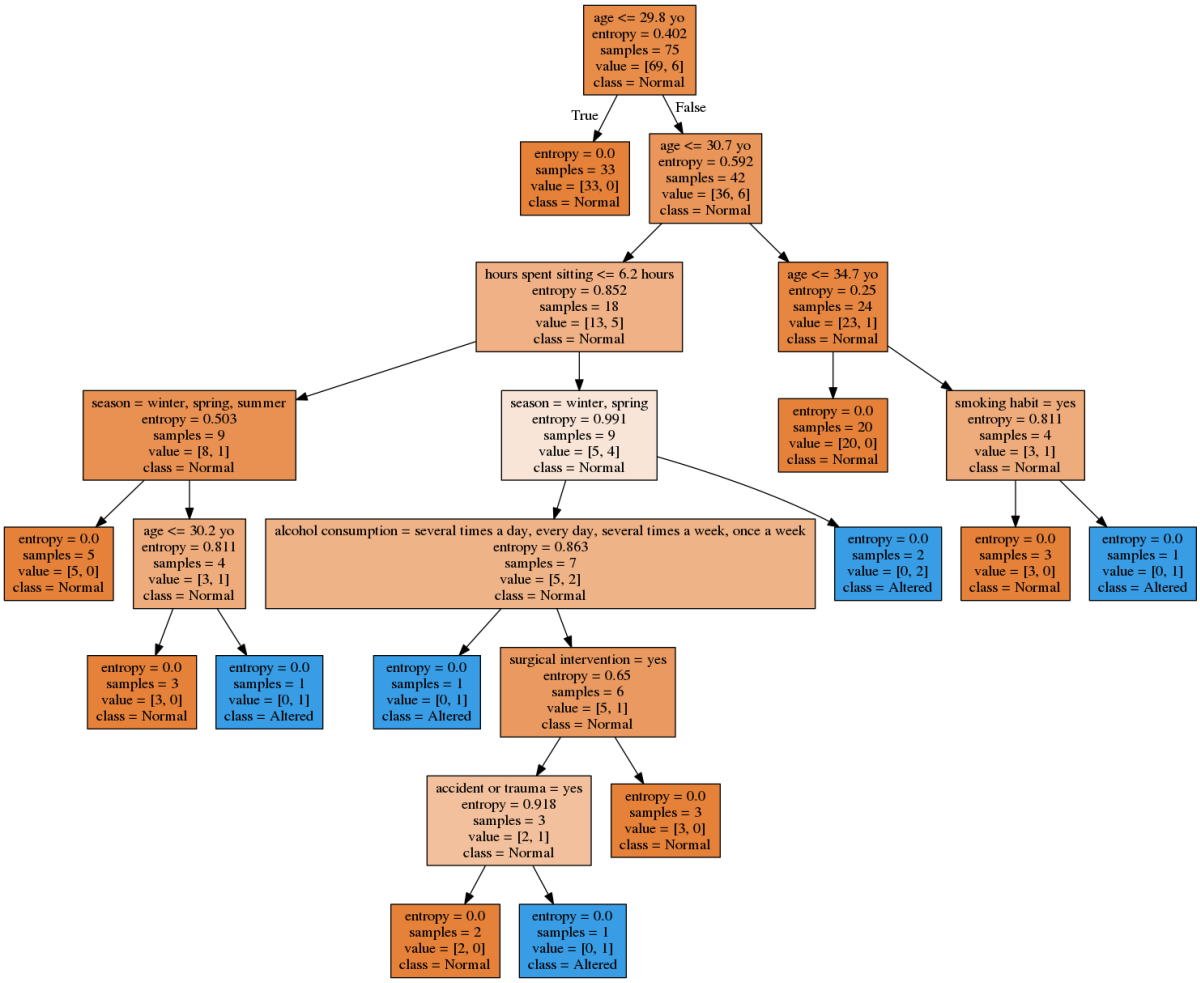
Tabel 3. *Confusion matrix*

Kelas Aktual	Kelas Prediksi	
	Normal	Altered
Normal	21	4
Altered	0	0

Model yang telah dibuat dapat divisualisasikan ke dalam pohon keputusan. Gambar 2 merupakan bentuk pohon keputusan yang terbentuk dari *data training* yang telah diolah. Dari model pohon keputusan ditemukan atribut-atribut yang menentukan tingkat fertilitas pria. Atribut yang menentukan seperti umur, lama duduk perhari, musim, kebiasaan merokok, konsumsi alkohol, mengalami pembedahan dan trauma serius. Atribut yang menjadi root adalah umur. Namun, terdapat atribut yang tidak digunakan, yaitu atribut Demam tinggi pada 1 tahun terakhir dan penyakit ketika kanak-kanak, karena tidak adanya hasil klasifikasi yang dipengaruhi secara signifikan oleh atribut-atribut tersebut.

Selanjutnya dari pohon keputusan pada Gambar 2 dapat dibentuk aturan atau pengetahuan, pohon keputusan merupakan himpunan aturan IF...THEN. (Romansyah, Sitanggang, & Nurdiati, 2009). Pada pohon

keputusan di setiap percabangan menyatakan kondisi yang harus dipenuhi dan tiap ujung pohon menyatakan nilai kelas data. Atribut Umur menjadi simpul paling atas dari pohon klasifikasi.



Gambar 2. Model Pohon Keputusan

Berdasarkan pohon keputuan di atas, dapat dibuat aturan sebagai berikut:

R1: IF (age <= 29.8) THEN diagnosis = normal

R2: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting <= 6.2) ^ (season = winter) v (season = spring) v (season = summer) THEN diagnosis = normal

R3: IF (age > 29.8) ^ (age <= 30.2) ^ (hours spent sitting <= 6.2) ^ (season = fall) THEN diagnosis = normal

R4: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting <= 6.2) ^ (season = fall) ^ (age > 30.2) THEN diagnosis = altered

R5: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting > 6.2) ^ (season = winter) v (season= spring) ^ (alcohol consumption = several times a day) v (alcohol consumption = every day) v (alcohol consumption = several

times a week) v (alcohol consumption = once a week) THEN diagnosis = altered

R6: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting > 6.2) ^ (season = winter) v (season = spring) ^ (alcohol consumption = hardly ever) v (alcohol consumption = never) ^ (surgical intervention = yes) ^ (accident or trauma = yes) THEN diagnosis = normal

R7: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting > 6.2) ^ (season = winter) v (season = spring) ^ (alcohol consumption = hardly ever) v (alcohol consumption = never) ^ (surgical intervention = yes) ^ (accident or trauma = no) THEN diagnosis = altered

R8: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting > 6.2) ^ (season = winter) v (season = spring) ^ (alcohol consumption = hardly ever) v (alcohol consumption = never) ^ (surgical intervention = no) THEN diagnosis = normal

R9: IF (age > 29.8) ^ (age <= 30.7) ^ (hours spent sitting > 6.2) ^ (season = summer) v (season = fall) THEN diagnosis = altered

R10: IF (age > 30.7) ^ (age <= 34.7) THEN diagnosis = normal

R11: IF (age > 34.7) ^ (smoking habit = yes) THEN diagnosis = normal

R12: IF (age > 34.7) ^ (smoking habit = no) THEN diagnosis = altered

Dari pohon keputusan dapat dilihat bahwa faktor utama yang dapat mempengaruhi tingkat fertilitas pria adalah umur dan duduk terlalu lama, sehingga dapat kita pedomani untuk merubah pola hidup agar tidak terus menerus duduk dan bekerja.

Ketiadaan data serupa di Indonesia membuat penulis hanya dapat menggunakan data yang tersedia di repository (http://archive.ics.uci.edu/ml/datasets/Fertility#) yang merupakan data sampel sperma 100 sukarelawan mahasiswa Universitas Alicante Spanyol. Untuk bisa diaplikasikan di Indonesia butuh penelaahan lanjutan dengan menggunakan data sampel sperma pria suku bangsa Indonesia berusia 18-36 tahun.

KESIMPULAN

Dari penelitian yang telah dilakukan, dapat disimpulkan bahwa algoritma classification and regression trees (CART) mampu menghasilkan pohon keputusan dengan tingkat akurasi sebesar 84%. Diharapkan model pohon keputusan yang dihasilkan dalam penelitian ini dapat digunakan untuk membantu dokter dalam prediksi tingkat fertilitas pria.

REFERENSI

Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M. (2013). Decision tree analysis on j48 algorithm for data mining. *Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*.

Susanto, B. M. (2013). Komparasi Algoritma Naive Bayes Dan C4. 5 Dalam Mendeteksi Kesuburan. *SNIT 2013*, 1(1), 69-73.

Foster, K. R., Koprowski, R., & Skufca, J. D. (2014). Machine learning, medical diagnosis, and biomedical engineering research - commentary. *BioMedical Engineering Online*. https://doi.org/10.1186/1475-925X-13-94

Gil, D., Girela, J. L., De Juan, J., Gomez-Torres, M. J., & Johnsson, M. (2012). Predicting seminal quality with artificial intelligence methods. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2012.05.028

Giwercman, A., & Giwercman, Y. L. (2011). Environmental factors and testicular function. *Best Practice and Research: Clinical Endocrinology and Metabolism*. https://doi.org/10.1016/j.beem.2010.09.011

Han, J., Kamber, M., & Pei, J. (2012). Data Mining. In *Data Mining: Concepts and Techniques*. https://doi.org/10.1016/C2009-0-61819-5

Irawan, G. A. (2017). Prediksi Kesuburan (Fertility) Dengan Menggunakan Principal Component Analysis Dan Klasifikasi Naive Bayes. *Jurnal Ilmiah Ilmu Komputer*. https://doi.org/10.24843/jik.2017.v10.i02.p02

Krausz, C. (2011). Male infertility: Pathogenesis and clinical diagnosis. *Best Practice and Research: Clinical Endocrinology and Metabolism*. https://doi.org/10.1016/j.beem.2010.08.006

Romansyah, F., Sitanggang, I. S., & Nurdiati, S. (2009). Fuzzy decision tree dengan algoritme ID3 pada data diabetes. *Internetworking Indonesia Journal*.

Rutkowski, L., Jaworski, M., Pietruczuk, L., & Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences*, *266*, 1–15. https://doi.org/10.1016/j.ins.2013.12.060

WHO. (2016). WHO | Infertility definitions and terminology. *Human Reproduction Programme*. https://doi.org/http://www.who.int/reproductivehealth/topics/infertility/definitions/en/